ETSI

ETSI TECHNICAL REPORT

ETR 261-4

October 1996

Source: ETSI TC-HF

Reference: DTR/HF-01028-4

ICS: 33.020

Key words: Keypad, MMI, supplementary service

Human Factors (HF); Assessment and definition of a harmonized minimum man-machine interface (MMI) for accessing and controlling public network based supplementary services; Part 4: Experimental comparison of the effect of categorized and non-categorized formats within user instructions

ETSI

European Telecommunications Standards Institute

ETSI Secretariat

Postal address: F-06921 Sophia Antipolis CEDEX - FRANCE
Office address: 650 Route des Lucioles - Sophia Antipolis - Valbonne - FRANCE
X.400: c=fr, a=atlas, p=etsi, s=secretariat - **Internet:** secretariat@etsi.fr

Tel.: +33 4 92 94 42 00 - Fax: +33 4 93 65 47 16

Contents

Blank page

Foreword

This ETSI Technical Report (ETR) has been produced by the Human Factors (HF) Technical Committee of the European Telecommunications Standards Institute (ETSI).

ETRs are informative documents resulting from ETSI studies which are not appropriate for European Telecommunication Standard (ETS) or Interim European Telecommunication Standard (I-ETS) status. An ETR may be used to publish material which is either of an informative nature, relating to the use or the application of ETSs or I-ETSs, or which is immature and not yet suitable for formal adoption as an ETS or an I-ETS.

Introduction

The Technical Committee for Human Factors has prepared this ETSI Technical Report to report publicly its work on the assessment and definition of a harmonized minimum man-machine interface for the access and control of public network based supplementary services. It is intended to complement ETS 300 738 [8].

This ETR constitutes part 4 of a multi-part ETR ("Assessment and definition of a harmonized minimum Man-Machine Interface (MMI) for accessing and controlling public network based supplementary services"), whose parts have the following titles:

Part 1: "General approach and summary of findings";

Part 2: "Literature review - Memory and related issues for dialling supplementary services using number codes";

Part 3: "Experimental comparison of two MMIs - Simulated UPT access and prototype ISDN supplementary services";

Part 4: "Experimental comparison of the effect of categorized and non-categorized formats within user instructions";

Part 5: "Experimental comparison of the CEPT and GSM codes schemes";

Part 6: "Survey of existing PSTN, ISDN and mobile networks, and a user survey of supplementary service use within Centrex and PBX environments";

Part 7: "Experimental evaluation of draft ETS 300 738".

Blank page

1 Scope

This multi-part ETSI Technical Report (ETR) presents the results of the research work conducted to develop a European Telecommunication Standard (ETS) defining a harmonized minimum man-machine interface (MMI) for the access and control of public network based telecommunications services, and in particular supplementary services.

This part 4 of the ETR describes the experimental comparison of two forms of user instruction manual, a categorized form and a non-categorized form. The experiment considered user performance and preferences within the basic user procedures provided to access a sample set of supplementary services. The experiment explored differences between a user manual structured to support and reflect a user model of supplementary services (categorized) and a user manual which is structured to reflect the necessary user procedures (non-categorized). The experiment also allowed a comparison between the performance of younger and older subjects.

2 References

For the purposes of this ETR, the following references apply:

[1] Bennett RW & Klinger JG (1990): "Conceptual models of telephony and their implications for interface design." 13th International Symposium on Human Factors in Telecommunications, Torino.

[2] Bouma H & Graafmans JAM (1992): "Gerontechnology." Amsterdam: IOS Press.

[3] CEPT T/CAC 02: "Subscriber control procedures for supplementary services in modern telecommunication system".

[4] Dooling JD & Klemmer ET (1992): "New technology for business telephone users: some findings from human factors studies." In RA Kasschau, R Lachman & KR Laughery (Eds.), Houston Symposium 3: "Information technology and psychology: prospects for the future." USA: Praeger.

[5] Dufour IG, Marshall JF & Welsby F (1991): "ISDN for the 1990s." British Telecommunications Engineering, 9, pp 246-252.

[6] Egly DG, Jeffries R, Leban B, Loebner EE, Parker L & Sears SB, (1985): "Mnemonic Aids for Telephone-Based Interfaces." 11th International Symposium on Human Factors in Telecommunications. Cesson Sevigne, France.

[7] ETS 300 511: "European digital cellular telecommunications system (Phase 2); Man Machine interface (MMI) of the mobile station (MS) (GSM 02.30)".

[8] ETS 300 738: "Human Factors (HF); Minimum Man Machine Interface (MMI) to public network based supplementary services".

[9] ETR 261-2: "Human Factors (HF); Assessment and definition of a harmonized minimum man-machine interface (MMI) for accessing and controlling public network based supplementary services; Part 2: Literature review - Memory and related issues for dialling supplementary services using number codes ".

[10] Fischer G (1991): "The importance of models in making complex systems comprehensible." In MJ Tauber & D Ackerman (Eds.) "Mental models and Human Computer Interaction 2." Amsterdam: North-Holland.

[11] Frankhuizen JL (1983): "Experiments with telephone user facilities." 10th International Symposium on Human Factors in Telecommunications. Helsinki, Finland.

[12] Holland JH, Holyoak KJ, Nisbett RE & Thagard PR (1987): "Induction: Processes of inference, learning and discovery." Cambridge MA: MIT Press.

[13] Israelski E (1988): "An experimental comparison of user performance with alternative access codes for PBX features." 12th Symposium on Human Factors in Telecommunications, The Hague.

[14] ITU-T Recommendation E.131: "Subscriber control procedures for supplementary telephone services".

[15] ITU-T Recommendation E.161: "Arrangement of figures, letters and symbols on telephones and other devices that can be used for gaining access to a telephone network".

[16] ITU-T Recommendation E.184: "Indications to users of ISDN terminals".

[17] Jones MLR (1990): "Making numeric command languages more usable." 13th Symposium on Human Factors in Telecommunications, Torino, pp 99-106.

[18] Judge P (1991): "Guide to IT Standards Makers and their Standards." London: Technology Appraisal.

[19] Kieras DE & Bovair S (1984): "The role of a mental model in learning to operate a device." Cognitive Science, 8, pp 255-273.

[20] Kieras DE & Polson PG (1985): "An approach to the formal analysis of user complexity." International Journal of Man Machine Studies, 22, pp 365-394.

[21] Marr D (1982): "Vision." San Francisco: Freeman.

[22] Neumeier J (1990): "A new user interface for ISDN telephones: Concept and prototypes." 13th International Symposium on Human Factors in Telecommunications, Torino, Italy.

[23] Nielsen J (1990): "A meta model for interacting with computers". Interacting with computers, 2, (2), 147-160

[24] Raven JC, Court JH & Raven J (1992): "Manual for Raven's progressive matrices and vocabulary scales. Section 3: Standard progressive matrices ", Oxford: Oxford Psychologists Press Ltd.

[25] Roberts TL & Engelbeck G (1989): "The effects of device technology on the usability of advanced telephone functions." Proceedings of CHI'89, New York: Association for Computing Machinery.

[26] Root RW & Koster CR (1986): "Experimental Evaluation of a Mnemonic Command Syntax for Controlling Advanced Telecommunications Services." Proceedings of the Human Factors Society 30th Annual Meeting, Santa Monica: The Human Factors Society.

[27] Rupietta W (1990): "Mental models and the design of user manuals". In MJ Tauber & D Ackerman (Eds.) "Mental models and Human Computer Interaction 2", Amsterdam: North-Holland.

[28] Rybash JM, Hoyer WJ & Roodin PA (1986): "Adult cognition and ageing: Developmental changes in processing, knowing and thinking." Frankfurt/Oxford: Pergamon Press.

[29] Schumacher RM (1992): "Phone-based interfaces: Research and guidelines." Proceedings of the Human Factors Society 36th Annual Meeting, Santa Monica CA: The Human Factors Society.

[30] Salthouse TA (1985): "A theory of cognitive ageing." Amsterdam: North-Holland.

[31] Veer GC van der (1990): "Human Computer Interaction: Learning, individual differences, and design recommendations." (Ph.D dissertation), Free University of Amsterdam, The Netherlands.

[32] Waern Y (1989): "Cognitive aspects of computer supported tasks." New York: John Wiley & Sons.

[33] Wickens CD (1990): "Engineering Psychology and Human Performance." Toronto: Charles E. Merill Publishing Company.

[34] Truijens CL (1985): "Symbols for supplementary telephone services: Experiments within CCITT." 11th International Symposium on Human Factors in Telecommunications, Cesson Sevigne, France.

[35] Zijlstra F & Meyman T (1989): "Het meten van mentale inspanning met behulp van een subjectieve methode." In T. Meijman (Ed.) "Mentale belasting en werkstress", Assen: Van Gorcum.

[36] ETR 261-1: "Human Factors (HF); Assessment and definition of a harmonized minimum man-machine interface (MMI) for accessing and controlling public network based supplementary services; Part 1: General approach and summary of findings".

3 Definitions, symbols and abbreviations

For the purposes of this part of the ETR, the definitions, symbols and abbreviations given in part 1 [36] of the ETR apply.

4 Theoretical Background

This clause consists of four subclauses. In 4.1 the usability of already existing supplementary service-like features is examined. Next, a theoretical framework is presented which serves to explain some of the observed phenomena (see subclause 4.2). Thereafter, specific aspects of elderly people using supplementary services are discussed in subclause 4.3. Finally, a number of research questions and hypothesis are formulated in subclause 4.4. It should be noted that experimental research into usability of phone based interfaces is limited and much of the work is unpublished (Schumacher, 1992) [29]. For that reason the current study must be regarded as an explorative one.

4.1 Usability of Existing Features and Functions

Modern private branch exchange systems (PBX; private telephony networks) already offer many of the features soon to be covered by ISDN supplementary services. However, several authors have pointed to the fact that most subscribers do not make use of the wide-ranging functionality of their terminals (Bennett & Klinger, 1990 [1]; Frankhuizen, 1983 [11]; Jones, 1990 [17]; Neumeier, 1990 [22]; Roberts & Engelbeck, 1989 [25]; Root & Koster, 1986 [26]; Truijens, 1985 [34]). There are a number of possible causes for this.

The most straightforward explanation is that users do not need the functions. The perceived costs of learning to use a service (too much effort reading a manual or exploring the device, loss of time) simply do not counterbalance the benefits (Dooling & Klemmer,1982 [4]). Whether users are conservative in this respect or whether the functions are really not useful is not clear. This question might be answered by performing a task analysis of the behaviour of people who know how to operate the system and actually make use of all of its available functionality. Furthermore, a distinction should be made between private users and business users. It is easy to imagine that people in business make use of certain services (e.g. teleconferencing, call forwarding, automatic call-back) far more often than private callers. To sum up, in some circumstances users might indeed be reluctant to learn operating procedures.

On the other hand it is also possible that subscribers do not know that the services are available. Root & Koster (1986) [26] found that simply advertising services was enough to increase the use of a service. These two explanations have to do with marketing aspects of introducing new services. As will be shown

in the rest of this clause use of services is certainly not only a matter of marketing. There are a number of issues which have to do with the usability of services.

Neumeier (1990) [22] concluded that advanced functions are not adequately presented to the user. The functions are there, but the user has to search for them before he can use them. Typically when there is only a phone based interface available the user has to recall all commands, which begin with a * or a # followed by a series of digits. Neumeier goes on to propose a new user interface concept which follows certain general principles:

- restriction to a small number of useful features;
- reduction of the number of keys;
- guiding the user by clear visual feedback.

According to Neumeier the only solution to users infinitely searching for the desired function is reducing the number of available functions. In addition, users might be confused by a large number of keys. Instead of "one function-one key" there is the soft-key concept: the relation between functions and keys is context dependent. Finally, Neumeier considers small Liquid Crystal Displays (LCD) actually inappropriate to give visual feedback. A large-screen graphical interface would be preferable. There are some remarks to be made with respect to the first and last principle. Simply reducing the number of available functions is a kill or cure solution. Assuming that a large number of functions is in fact necessary, improving search strategies is another option. Moreover, constructing comprehensive texts on a small screen LCD is possible, but will require creative thinking on the part of user interface designers.

Jones (1990) [17], who conducted a more rigorous analysis, lists a number of difficulties of numeric command languages:

- lack of task structuring;
- lack of contextual information;
- lack of perceptual cues;
- demands on memory and problem solving abilities.

In the field of Human Computer Interaction (HCI) the solution to the problem of memorizing commands was found through the invention of menu-based and later on graphical dialogues on the screen. This meant that the commands could be displayed semi-permanent. However, phone based interfaces provide auditory "displays" only. Menu presentation is necessarily serial and may turn out to be lengthy in the case of an appreciable number of functions. Moreover, auditorily displayed menus may place heavy demands on short term memory, thereby limiting elderly people especially. Visual displays are more appropriate for the parallel presentation of menu-items (Schumacher, 1992 [29]). However, even PBI++ displays are still small (less than 40 characters). So, if it is decided to use visually displayed menus, the minimal interface may in fact turn out to be not so minimal after all.

Bennett and Klinger (1990) [1] conducted interviews with professional and managerial employees of companies that used several varieties of PBX. No one used an "appreciable" number of features unless their job required that they do so. (It is not clear what Bennett & Klinger call an appreciable number; from TC-HF members' informal observations with PBXs, it appears to be approximately three to five out of ten to fifteen.) After provision of an instruction booklet, they understood and readily followed the instructions. However, the actions they performed would always be arbitrary and mysterious to them. It appeared that most of them had little faith in their understanding of what was "really happening" when they pushed a button or dialled a code. There may be severe consequences for the interaction with a phone based interface if users feel that they do not "understand what is going on inside". A classic example is that of a user who has to transfer a call. In order to do so (on some PBXs) it is necessary to flash the switch-hook. This means pressing the "switch-hook" - which is normally used to terminate a call - for a very short while (200 to 1 100 ms). Most users consider this too closely corresponding to terminating a call. It has been found that only 40 % of attempts to use the switch-hook as a signal within a call are successful. A difficulty related to the comprehensibility of services procedures is the abstract nature of the services themselves. Users cannot imagine what kind of processes are going on inside their telephone or the exchange (Truijens, 1985 [34]). This issue will be elaborated fully in subclause 4.2.

4.1.1 Summary

By now a number of problems in existing phone based interfaces have been identified. The most important issues concern difficulty remembering the commands, lack of feedback on which action was

performed, and the resulting system status, and finally lack of understanding of what is happening "inside".

4.2 Mental models

If a user wants to engage in an interaction with any system - the system may in fact be 'the world', containing an infinite number of subsystems - he must have knowledge of it. The cognitive structure that constitutes this knowledge is called a *mental model* or *mental representation*. A mental representation is a simplified abstract representation of the real system. It allows people to infer what to do to put the system into a desired goal state and to predict what the consequences of a certain action will be.

The problems summarized in subclause 4.1.1 can be reworded in terms of mental representations in order to produce some new insights.

4.2.1 Mental models of supplementary services

In order to investigate user aspects of supplementary services in telephony, it might be useful to see first what kinds of mental representations people possess of (normal) telephony. However, little research has been done and therefore the argument is necessarily tentative for the most part.

4.2.1.1 Mental representations of telephony

An ordinary telephone can be viewed as a tool which is used to complete a certain task. This task might be "to convey a spoken message to another person who is not close enough to communicate with in a direct way". As said before, in order to be able to use the telephone it is necessary to have some kind of representation of its functioning. In this respect there is an important resemblance between telephones and computers: their functions are not apparent from their structures. (Mechanical devices may be complex and hard to understand, but it is always possible to ask an engineer to show the inside and explain the functioning of each physical component. This is not the case with information technology, because of the inherently abstract nature of the components (*information* and *processing*) themselves.) The absence of this *user-perceivable structure* makes it impossible for users to acquaint themselves with closed subparts of the system (Fischer, 1991 [10]). As a consequence, the mental models most people employ in telephony are very rudimentary. For instance, they hardly allow for the representation of network characteristics. Figure 1 shows the most stereotypical models Bennett & Klinger (1990) [1] encountered in their study of 26 experienced business telephone users' representations. Panel A shows the simplest model (8 subjects), assuming that each party in the call has a telephone and that there is a wire connecting the two, and little of anything else. People with a model like the one shown in panel B (3 subjects) assumed an arrangement similar to some experimental distributed packet systems. They believed there were some common facilities and that station sets were "smart" enough to send their voice to the other set(s) in the call. Finally, five subjects had models like that shown in panel C. They believed there was an "automated operator" that connected telephones together in response to dialled digits. They either had no idea (or interest) in how many automated operators there might be, or assumed there was one big one somewhere in the middle of the country. The remaining subjects lacked a discernible model, or they may have had a model that was not understood by the researchers.

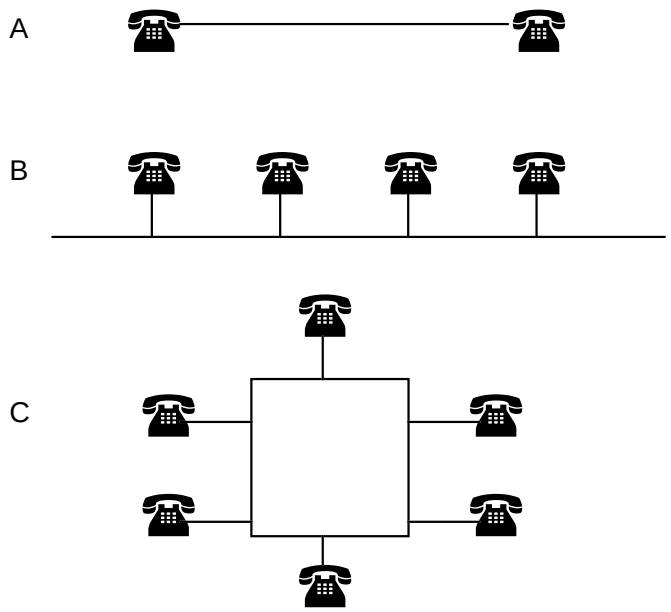

Figure 1: Mental models of telephony (after Bennett & Klinger, 1990 [1])

Note that Bennett & Klinger used relatively experienced subjects. It may be expected that most "normal" users have models like the one in panel A of figure 1 or none. Besides that, Bennett & Klinger noticed that most of the subjects had never given much thought to how telephony worked. So, the models they found were either used implicitly or the result of ad hoc reasoning.

As far as the presence and complexity of mental representations is concerned, there is a difference between telephony and the field of HCI. It has been found that people are in fact able to build quite complex and adequate mental models of the computer systems they work with (Van der Veer, 1990 [31]). This difference is explained by the fact that no one ever receives training in using a telephone, either formal or informal. Most people learn to operate it by experimentation at an early age, mimicking adults. They develop their own representations while exploring the device. This seems to be enough to operate it adequately, which is not so strange of course, since operating procedures are very simple. Kieras & Bovair (1984) [19] conclude that the telephone book contains enough "how-to-do-it" information, and that it is in fact unnecessary to supply "how-it-works-information". Moreover, the normal conversation might be a sufficient metaphor for ordinary telephony: if you want to talk to someone, you call his name (number) and you talk to him. If the person is not present he will not answer. If he is busy he will tell you so (busy tone). If something has gone wrong you are told . If you want to, you can try to call him at a later time, etc.

4.2.1.2 Mental representations of supplementary services

While increasing, network capabilities have lead to the creation of new services, the goal is still to use the ordinary telephone as a physical interface. One of the reasons is of course the omnipresence of the device throughout the world. It would be an enormous advantage if customers for a change would not have to purchase new equipment to acquire a service. Besides that, it is feared that the introduction of all kinds of extra buttons on new telephones may lead "ordinary" people to stay away from buying them.

A side effect is that people do not expect their telephones to be able to offer this functionality. The looks of the ordinary telephone trigger the "old" mental model that belongs to it. This model does not allow for any network concepts. Consequently, many people are very surprised when they are told that their telephone is capable of doing "all these thing" as well. This is explained by the fact that a great deal of the functionality of services is no longer in the telephone itself but in the network. However, it has been shown that the models people employ do not contain any representation of network elements (e.g. an exchange or a metaphor for it).

To most users it is not apparent what the facilities are, let alone their operating procedures. The transition from a situation where a very rudimentary mental model suffices for adequate behaviour to a new one where the (cognitive) user-interface is based on a complicated and necessarily abstract conceptual model seems to cause another problem. In other words, the old model of a wire connecting the two callers is insufficient to tell the user how to operate advanced services (figure 2).

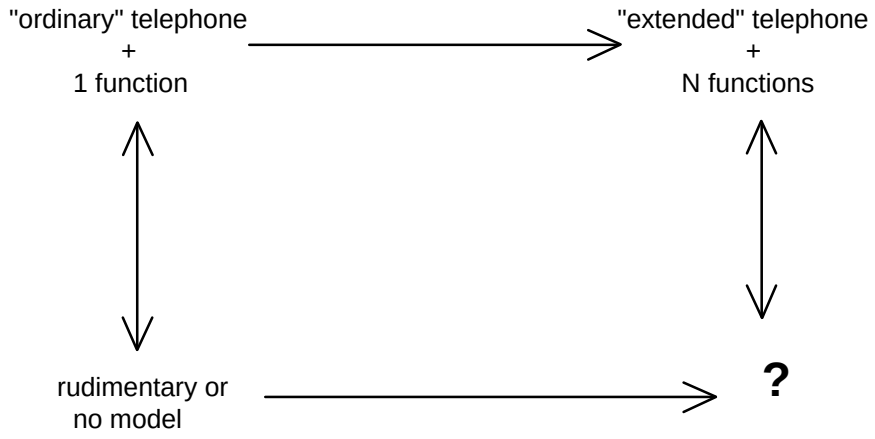

Figure 2: The transition from a system where a rudimentary model or none suffices for adequate behaviour to a new system which is based on a complex concept

Finding a solution to this problem is a major challenge but requires extensive research. Therefore it does not come within the scope of this study. However, while it *is* stressed that this should be a topic of major concern to service providers, a few remarks can be made.

Kieras & Bovair (1984) [19] have shown that supplying the right "device model" (i.e. the information presented to the user) can actually support learning how to use a system. They come up with the following suggestions:

- The device model information must support inferences about the exact and specific control actions. Thus, for example teaching users general principles, metaphors, or analogies will be of little value, since these are unlikely to support such precise inferences. On the other hand, these general principles etc. might aid the user in constructing his mental representation, as long as this (redundant) information does not hamper the interaction. Besides that, users may possess different mental representations of the device, depending on the task that is being carried out. These representations may be connected by some kind of general principle, metaphor or analogy. The central point here is that the information which is supplied minimally allows for the above-mentioned inferences.

- The relevant "how-it-works" knowledge can be very superficial and incomplete, because the user does not need to have a full understanding of the system in order to be able to infer the procedures for operating it.

4.2.1.3 Remembering the commands

It is not necessary to understand how the services *really* work; it is enough to have a mental representation which allows for the inference of the commands to issue. Assuming, therefore, that people in fact develop representations which allow them to understand the functionality of these kinds of service, another issue is remembering what actions to perform to operate them. In order to (de-)activate a service, the user must establish two links: first, between the goal to be accomplished (establishing a certain state in the system) and the specific command or commands that accomplish that goal, and second, between the command (e.g. forward all incoming calls) and the key presses that initiate that command (Egly et al., 1985 [6]). Available research shows that some improvements are possible which might facilitate the user in remembering the relation between commands and key presses.

In current phone based interfaces, the procedure to (de-)activate supplementary-service-like functions consists in typing *, # and digit keys. To the user this type of command seems to be an arbitrary string with no inherent meaning or relation to the effect of typing it. Users of PBXs may sometimes have to learn 40 or more different commands if they want to employ all of its functionality. Lacking a coherent framework within which to organize the services command set, users must resort to rote memorization.

Several authors have pointed to the fact that a mnemonic command syntax may aid the user in remembering the commands (Egly et al., 1985 [6]; Israelski, 1988 [13]; Jones, 1990 [17]; Root & Koster, 1986 [26]). Israelski (1988) [13] did a series of laboratory experiments involving 178 subjects in order to determine an optimal set of dial code plans for accessing PBX voice features. The four code plans involved 1, 2 or 3 digits with and without the special symbols * and #. In the experiments, subjects were trained as normal users and were asked to play the role of business office PBX users. The role playing required subjects to solve twenty problems in each of two separate sessions by using special PBX voice features such as conference, transfer, hold and automatic call back. Results showed performance and preference superiority of a two character code plan using * and # symbols in conjunction with a *single* digit. Israelski suggests that less frequently used feature access codes could have a two digit structure preceded by a * or a #. However, this is not logical, because service commands that are not used frequently are hard to remember already. Giving them a two digit code would only worsen that.

Root & Koster (1986) [26] designed a command set that featured an explicit syntactic structure, mnemonic service-control codes, user-assignable strings for personal dialling lists, and service-independent codes for functions such as service cancellation. They also conducted an experimental evaluation of the syntax and found that users generally remembered more services and were more accurate at recalling service commands than were users of the "old" numeric interface (which had no structure).

Finally, Jones (1990) [17] formulated three guidelines for designing commands which take into account people's natural learning strategies:

Presentation: Commands should be explicitly presented as consisting of groups of functions rather than an unstructured list of items.

Relation: There must be an identifiable, non-arbitrary relation between the members of a group.

Group Digit: All codes for functions within a group, should begin with a common group digit to reinforce the grouping of functions.

Jones' aim was to enable a user to form a conceptual model of the language as a whole. The guideline which refers to *Presentation* concerns consistency of the syntax as well. Jones also cites authors who have found that non-numeric delimiters can be used to great effect to make the syntactic structure clear and thus improve usability. The *Relation* guideline is based on the notion that people can remember much larger amounts of information if that information is presented and learnt as a structure consisting of related "chunks" or categories, rather than a list of unstructured items. Jones remarks rightly that the groupings need not reflect the true physical structure of the domain to which it applies. What counts is the (user) model's *functional* role in representing the world. Furthermore, chunking of commands into groups should be based on function similarity. Exactly how functions in a group are similar to each other may vary from group to group: in one group all the functions may achieve a similar *effect* (e.g. "call forwarding"). In another, the functions may all relate to a particular *event* (e.g. "if you want to call more than one person"). Depending on the availability of external prompting, the maximum number of groups is 7 or more. Moreover, the maximum number of items *within* a group should not exceed 7. An important usability issue is the fact that, according to Jones, functional grouping seems to be of particular help to intermittent knowledgeable and novice users. Finally, the *Group Digit* guideline states that where pairs of functions exist which refer to a common facility (e.g. switching a function on and off), usability can be improved by assigning congruent codes to them (Jones, 1990 [17]).

On the basis of these three guidelines Jones constructed an artificial PBX-like command language for instructing a robot to do household tasks. In a subsequent experiment, subjects were instructed to remember as many as they could of 47 commands in 10 minutes. It was found that the *combination* of the above-mentioned characteristics had a strongly significant influence on recall of service codes. By choosing a familiar domain, she avoided the problem of users' inability to understand the functionality of services (see above), resulting in a "purer" measure of recall of familiar commands. Although this is an advantage on the one hand, a few remarks should be made. Jones' subjects would not have to engage in problem solving activities. However, it has been found that problem solving processes are a major source of mental model formation and modification (Holland et al., 1987 [12]). Besides, difficulties associated with

the comprehensibility of function names and of function actions may be expected to inhibit the successful construction of a coherent mental model. With regard to the comprehensibility of the actions, Jones points to the fact that if a user is simply unfamiliar with the *concepts* (see previous subclause) then the power of functional groupings will be limited. A final yet very important issue is the learning environment. The incentives present in an experiment (being paid, and knowing a test will be given), differ from those in real usage (in business telephony getting a transfer wrong could mean losing a customer). Another problem is the type of learning. In the experiment, subjects were instructed to remember as many codes as they could. In "real life" a user will start by looking up the codes whenever they are needed. This means that learning is merely incidental and not intentional. That is, a user may try to learn one or several codes intentionally, but certainly not *all* the codes.

4.2.1.4 Information presented by the system

Learning processes and feedback are closely intertwined. Considering the fact that the majority of current telephone users are unacquainted with phone based interfaces in general, a growing number of researchers have become interested in the way people (learn to) operate them. Moreover, as it can be argued that even experts are constantly refining their mental representations (Holland et al., 1987 [12]) one must conclude that feedback is an important issue.

Information presented by any system should consist of three types: first of all, the actions a user performs should be acknowledged by the system. This way the user can decide whether the system received the command he intended it to get. Making the action response relation explicit is especially important in the case of novice users, in order to facilitate the building of a correct mental representation (and avoid the building of an erroneous one!). Besides, a first time user seeing that he only made a keying error might be comforted instead of scared off. His self-confidence might even grow as he discovers that in case of a keying error, the system's reaction indeed corresponds to the way he would expect it to (i.e. according to his mental representation of it). Secondly, a user should be informed of the system status that resulted from the actions he performed. This way the user always knows "where he is", which makes navigation through the interface easier and places less demands on short term memory (STM). Thirdly, in any stage of the interaction the user should be made aware - "in a context sensitive way" - of the commands that are allowed at that time and their effects (the user is prompted by the system). The first and second types of information usually coincide, but this is not always the case.

Looking at existing supplementary service-like features in PBXs it must be concluded that neither of the above-mentioned criteria is entirely satisfied. Although the touch-tone keypad does provide some feedback by means of DTMF tones, it is not realistic to expect users to infer which digit they just dialled without doing a lot of conscious processing. System status is usually presented by means of single tones. They bear a number of important limitations. For instance, a person's ability to learn, distinguish between, and remember, different tones representing abstract conditions is limited to about four to six. Users are frequently confused by unfamiliar tones encountered through travel or international communication. For this reason ITU-T Recommendation E.184 [16] advised against the use of new tones. The use of tones requests a substantial amount of a priori knowledge regarding their meanings. Taking into account the poor mental models users have of telephony, it is indeed not likely that they will soon learn the meaning of new tones. They just do not fit within the mental representation. More informative feedback might facilitate the building of the model. In other words, feedback should "teach" the user, instead of expect him to know already. With regard to the third criterion, the user is mostly prompted by the dialling tone only at the beginning of a call or service (de-)activation procedure. Ideally, a (novice) user should be guided throughout the entire procedure, receiving feedback after each keying action. Apart from the issues mentioned above, it must be said that until recently, technical limitations have been an important factor. It was simply impossible to develop more sophisticated phone based interfaces. With the advent of ISDN as a replacement or extension of the existing PSTN, it is to be hoped that the situation will change.

4.2.2 Presentation: how to communicate the conceptual model to the user

If mental representations are really that important, how then could one be sure that novice users develop the right ones? To answer this question it is necessary to make a distinction between the various types of model. In the field of HCI, Nielsen (1990) [23] constructed a taxonomy of all the various types of model that designers, (computer) systems, users and researchers may have of each other, (cognitive) tasks, manuals and the surrounding world. Only the ones relevant to the present study will be dealt with.

4.2.2.1 The (designer's) intended conceptual model, the (manual's) transmitted conceptual model and the (user's) actual conceptual model

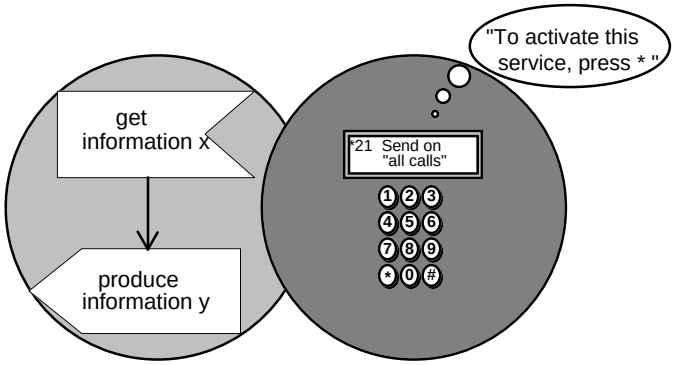

Figure 3: The designer's model (left) and the user's model (right) of supplementary service on a phone based interface

In developing a new system, a designer must have some kind of idea about its functional capabilities and information flows. Usually this *designer's* model is externalized, by means of diagrams, specifications, etc. Very often too, details of the system's design are left implicit and internalized and do not become concrete until the implementation stage. This holds for the user interface as well, or perhaps even especially. Anyway, the designer has to have some kind of idea about how the system should look and function and also about how a user ought to think of it and operate it (figure 3).

The *transmitted* conceptual model is the "manual's model of the system" (see also Rupietta, 1990 [27]). It is the version of the designer's model that makes it to the manual. Although Nielsen considers this the only way the designer's model is transmitted, it can be argued that the *user's actual conceptual model* depends on a much more heterogeneous set of information.

To start with, the system itself may be a source of knowledge (Rupietta, 1990 [27]). Through the provision of clear and informative feedback it should be possible to aid the user in constructing his mental model. Secondly, there are fellow users who have more (possibly erroneous) understanding than the user himself. Many people consult other users in order to fill the gaps in their knowledge. A third way users infer operating procedures of a new device is by transfer of knowledge from one domain to another. In this regard, Kieras & Bovair (1984) [19] make a distinction between device-dependent and device-*in*dependent knowledge. The ease with which a user acquires knowledge of a new system depends on the amount of device-dependent knowledge he has to incorporate. Supplementary services seem to present a somewhat deviant case, because on the outside the operation of a telephone appears to remain fairly constant (to dial a (code) number requires device-*in*dependent knowledge). However, the mental representation that is needed to infer what to do is changed dramatically (see previous subclause). Besides, for a user transiting from ordinary telephony to telephony including supplementary services, there is an appreciable amount of device-dependent knowledge to gather about the command language that is used to operate the supplementary services. The main focus of ETS 300 738 [8] is to develop a standard which ensures that for a user who transfers *between* technologies A and B, for instance GSM and UPT, or countries, knowledge about the user interface language is device-*in*dependent (figure 4). This way, a user is not bothered by having to learn new operating procedures.

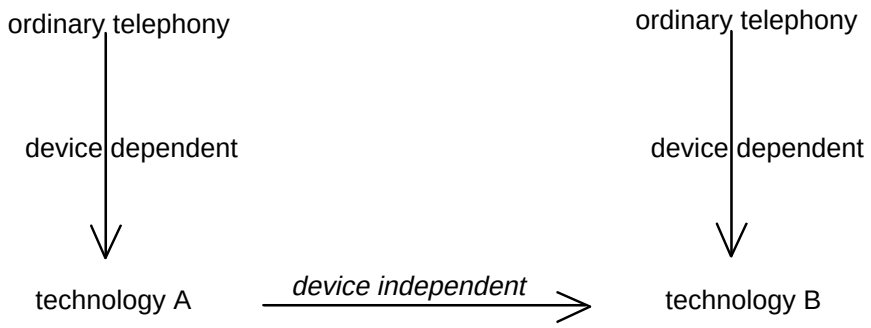

Figure 4: Acquisition of device-dependent knowledge and transfer of device-independent knowledge

In the aforementioned cases, mental model formation proceeds in a rather unstructured way, and one is not sure that users develop the "right" representations. Tutorial facilities outside or within the system are the formal media through which the user ought to be taught how to operate the system. This way, building the representation is controlled to a certain extent, so that the user might be expected to develop a model that corresponds quite closely to the intended conceptual model. However, one is never sure that a novice in fact develops the "right" presentation, or as Nielsen [23] formulates it: "... *there is no guarantee that the internalized model formed by a reader of an externalized model is that intended by the writer of the model ... The important point is to remember that there is a difference between what we intend users to think (or what we postulate that they think) and what they actually think.*" The chance of a discrepancy between the three models is probably dependent on the complexity of the first and second one. In other words, the transmitted conceptual model should be as simple as possible. Redundant information is allowed as long as it does not add to the complexity of the model.

4.2.2.2 The information retrieval model

Finally, there is the user's model of where to find information about the system: the *information retrieval model*. This model of the system and manual is in some way a sub-model of the user's functional mental model, but it may be useful to view it as a separate model as it also includes the user's model of the manual and as there is some difference between what the user knows about the system and what the user knows about how to learn more about the system (Nielsen, 1990 [23]). As a manual may be organized according to the structure of the interface language (e.g. the order and names of the sections may be the same as that of the items in the main menu, see clause 5), the user's mental representation of both may be quite similar. It should be noted that nowadays information retrieval models (for computers) are becoming more and more device-independent. In other words, many computer software applications contain on-line context sensitive help functions that are accessed in a progressively more consistent way.

As a concluding remark, it is stated once more that the designer's model and that of the user need by no means to be the same. In the end it is not the exact nature of the user's model that is important, but its *functional role* in allowing the user to operate the system. It is the responsibility of the designer (of the user manual) to transmit an easy to incorporate and "action facilitating" model. Clause 5 describes two different transmitted conceptual models of an experimental command language for accessing ISDN supplementary services.

4.3 Elderly users

ISDN supplementary services will in due course become available on the public network. As a consequence, the intended audience will be a very heterogeneous one. One of the most rapidly growing subgroups of telephone users consists of elderly people whose level of computer experience is little to none. On the other hand, in case of decreasing physical mobility, they may turn out to be the most frequent and therefore important users of telecommunication and phone based services in general. However, this will happen *only if services are useful and usable to them.* For this reason it is necessary to examine their ways of operating phone based services. Although ISDN supplementary services are themselves of interest, it is pointed out here that they might be taken as a "worst case" representative of other phone based services. If elderly users are able to operate them in an easy way, they will surely be able to operate other phone based services. On the other hand, if the interaction does not go so smoothly, it is not necessarily so that the results may be generalized to other phone based interfaces.

Supplementary services are more difficult to "understand", because they involve unknown abstract network concepts, as opposed to other phone based services for which usually some kind of practical daily life "shopping" metaphor might be used.

4.3.1 Elderly people's models of devices

Researchers of elderly people's interaction with information technology commonly agree that, for certain kinds of device, elderly people experience more problems than younger people. In fact, a new field has emerged which deals with this topic specifically (Bouma & Graafmans, 1992 [2]). In this subclause it is argued that the nature of the problems which elderly users of information technology encounter may be best explained by employing the notion of mental representations once again.

Effects of ageing can be roughly divided in two broad categories. On the one hand a number of neuro-physiological abilities are found to decline with age. In general, less and less complex information can be processed per unit of time in a less precise manner. For example it has been shown that elderly people's performance on measures involving (division of) attention, vision, integration of information, inductive abilities and (short term) memory is significantly impaired if compared to younger people (Rybash et al., 1986 [28]; Salthouse, 1985 [30]). On the other hand, elderly people have social, educational, professional and economic backgrounds which by nature differ widely from those of relatively young people. Consequently, their views of all kinds of devices may differ from younger people's as well. With respect to the (in-)ability to use information technology, age-related differences belonging to the latter category are probably much more important than it would seem at first sight. Although it *is* recognized that those from the former do play a role, along with - for instance - motivational issues, the notion of a mental representation also serves to explain why elderly people have trouble using information technology.

During past decades there has been a transition from the Industrial to the Information "Age". From a cognitive point of view this might be seen as the change from the *mechanical* to the *information processing* society. Until halfway through this century, device functions were usually quite simple and fairly straightforward to use. In case of more complex machinery, users' ways of operating devices were based on "mechanical" representations. As mentioned before, mechanical devices may be complex and hard to understand too, but it was always possible to ask an engineer to show the inside and explain the functioning of each physical component or provide a concrete metaphor for it. Besides, as mentioned before, for a lot of mechanical devices it is only necessary to understand their external functionality, which can be perceived from the outside. For instance it is not necessary to be able to understand the internal functioning of a clock in order to be able to see what time it is, or to understand the internal working of a car to drive it. This *is* in fact the case with many "information devices", because of the inherently abstract nature of the components (information and processing) themselves and because the user has to manipulate abstract (data) structures *inside* the system. Most user interfaces do not allow for mechanical representations to explain how to use them, so elderly people have to learn to view "information devices" in another way. They have to get acquainted with new and more abstract concepts. However, ageing is known to influence the ease with which a person acquires new material. Moreover, the existing "mechanical point of view" may interfere with the new way of approaching devices. In other words, elderly people who have a lifetime experience of explaining device-behaviour using physical mechanical concepts may be hampered in particular when trying to grasp the new and abstract informational concepts used in current user-interfaces. If this is true, there is an important consequence for the demarcation of the group of users who have trouble using information technology. Age would not be the most important factor. Rather, familiarity with abstract (information processing) concepts determines whether a person is able to grasp the transmitted conceptual model of an "information device". Although it is recognized that there are a lot of confounding factors involved, the current experiment aimed at discovering the tenability of this hypothesis.

4.4 Hypothesis, expectations and research questions

The theoretical insights above lead to a number of expectations as well as questions:

4.4.1 Research questions

Are novice users able to form and use conceptual models from a very brief description of the services themselves? Do people learn in this respect, i.e. develop their models while working with the services?

These questions will be answered by examining performance changes of users trying to solve "telecommunications problems".

4.4.2 Hypothesis

1) Assuming that a user understands the functionality of the supplementary services involved, a command language with an explicitly offered structure will facilitate the user in remembering and using the commands.

   - How quick will people be able to grasp the conceptual model of the interface language?

   - Differences between elderly and younger subjects are expected. How large will these differences be?

   - Will any differences in mental representations (as a result of different transmitted models of the interface language) effect observable behaviour?

   These questions will be answered by comparing two groups of users, who either receive an explicit or an implicit conceptual model of the structure of an experimental command language. Within these groups a distinction is made between elderly and younger users.

2) Elderly people and younger people with no abstract experience and or abilities will have trouble using the supplementary services.

   - To what extent will they still be able to use them? Are "simple" services (functionality) easier to operate?

3) A relation between performance and experience with respect to the ability to handle abstract concepts is expected.

   - To what extent do professional occupation - e.g. a job that requires using computers, but also other abstract non-technical jobs - and education, influence performance?

The next clause deals with two versions of an experimental command language that were designed to test the effect of presenting an explicit model of language structure.

5 Two different conceptual models

In order to investigate the effect of presenting an explicit structure of a command language, two different types of conceptual model of an experimental command language were designed. It was decided that information about the user interface language should be communicated to the user by means of a user manual. Two different types of manuals were constructed. It should be pointed out that both contained exactly the same information about the *functionality* of the supplementary services that were used. There was only a difference in the *organization* of the material, i.e. in the transmitted conceptual model of the interface (command) language. With respect to the models that were transmitted a distinction was made between *comprehensive* "how-it-works" knowledge in the one manual and *procedural* "how-to-do-it" knowledge (Kieras & Polson, 1985 [20]) in the other. This was knowledge *with regard to the language* and not as to the functionality of the individual services.

To isolate the effect of the conceptual model of the language, it was decided that user manuals should contain no explicit conceptual model that explained the functionality of the services involved. It was left an open question whether users are able to form their own concepts. To avoid information overload the descriptions in the manuals were kept as simple as possible. For each service there was information about:

- its functionality (what happens on (de-)activating it);
- when to use it (at call set-up, during call, call clearing phase, etc.);
- how to operate it (differed between the manuals).

5.1 The categories manual

The categories manual (CAT) was constructed to communicate an explicit conceptual model of the user interface command language. It was intended to capitalize on the user's comprehension of its internal structure.

5.1.1 A narrow two layer hierarchical model

The language was structured according to two models: the model of the *service codes* and the model of the *syntax*.

5.1.1.1 Six categories

In the same way Jones (1990) [17] built her "robot command language" (see subclause 4.2.1), the supplementary services were divided into six categories. Each category was given a name that corresponded to the functional parts that the services within the category had in common. Next to a name, each service had a unique corresponding two digit code.

NOTE: CEPT has in fact made an effort to build a structured code scheme as well (T/CAC 02 [3]). However, it is less consistent (sometimes 3 digit codes) and many codes are reserved for national use. Where possible the experimental codes follow the CEPT scheme.

The first digit in the code was the same for all members of a group. It was called the "group number".

5.1.1.2 Services within categories

Within each category the services were given names as well. This name usually consisted of two parts. The first part referred to the group the service belonged to. The second part referred to an aspect of its functionality that distinguished it from the other group members. The second digit in the service code denoted the service within the group, the "service number".

Example: The group *"Send on"* had group number *2*, the service *"Send on - all calls"* had service number *1* and thus the code that corresponded to it was *2 1* (although it was never referred to in that way). This way of structuring the set of service codes was chosen in order to exploit people's natural ways of learning (hierarchical category formation; Holland et al., 1987 [12]).

5.1.1.3 Syntax

In order to encourage the user to acquire the conceptual model of the command language's syntax in an active way, the actual complete code string that should be entered to operate a supplementary service was nowhere available. Users had to construct the commands *themselves.* An important aspect of the syntax was its high degree of consistency in case of correctly typed commands. To activate a service, all delimiters should be *, to deactivate a service all should be #. Operating a service involved typing the service code within two delimiters. In case additional information was required (activation only) the command was extended by appending the extra information followed by a *. This "symmetric" structure should facilitate the user's recall of the syntax. Again the intention was to exploit people's natural ways of information processing (the - visual - cognitive system is highly susceptible to symmetric structures; Marr, 1982) [21].

Example: To activate CFU, the user types *21*123456*, to deactivate it #21#; to activate HOLD the user types *50*, to deactivate it #50#; to activate CLIP *10*, to activate CLIR #10#, etc.

5.1.2 Table of Services

Table 1 shows a typical selection from the CAT Table of Services.

NOTE: *Table of Services* and *User Manual* refer to the actual manuals representing the two conditions. The terms are used in conjunction with CAT or NoCAT. "*manual*" (no capital) refers either to user manuals in general or the combination of User Manual and Table of Services.

Table 1: A typical section from the CAT Table of Services

GROUP 2	**Send on**
Service 0	not in use
Service 1	**Send on - "all calls"** This service allows you to send on *everyone* who calls you up to another telephone automatically. For that purpose you must enter the number of that telephone (*the destination*), after activating the service. *How to use it?* Activation and deactivation is possible only if you hear the dialling tone.
Service 2	**Send on - " when busy"** This service allows you to send on *everyone* who calls you up to another telephone automatically. For that purpose you must enter the number of that telephone (*the destination*), after activating the service. Phone calls are sent on ONLY if your telephone is BUSY. *How to use it?* Activation and deactivation is possible only if you hear the dialling tone.

As a result of the fact that the actual code strings were not to be available to the user, it was possible to separate the conceptual model of the syntax from the service codes. For that purpose a separate Table of Services was constructed which contained service codes only. Again, this Table was organized in such a way that the conceptual model of the hierarchical group structure was reinforced. The first page was a list of the groups, containing group number, group name and a short description of the group members' shared functionality. The rest of the Table contained descriptions of individual services like the ones shown in table 1. Note that the conceptual model is reinforced by - firstly - forcing the user to infer the service code (21) from the group number (2) and the service number (1) and - secondly - showing unavailable service numbers as well.

5.1.3 User Manual

Apart from the Table of Services a User Manual was constructed that contained a short introduction on the supplementary services. Furthermore, it explained the group and syntax structure, and that there were group and service names. Finally, there was some general information about the user interface. The information on syntax was preceded by the summary shown in table 2. This summary may be viewed as the transmitted conceptual model of the syntax.

Table 2: The conceptual model of the syntax that was explained in the CAT User Manual

Summary	
to activate	* *group* nr. *service* nr. * (go back with #)
to *de*activate	# *group* nr. *service* nr. # (go back with *)
	* and # are each others' opposites

5.2 The No Categories manual

In order to set out performance in the former condition against a baseline, the NoCAT manual was supposed to reflect the current state of Dutch phone-book instructions for CFU. Moreover, it was meant to reflect a command language with an implicit structure.

5.2.1 A flat and broad model

Any user manual communicates some kind of model to the user. As said before, the contents of the CAT and NoCAT manuals did not differ, as far as service codes, service names, syntax and descriptions of services were concerned. Nevertheless, there *was* a difference between the CAT and NoCAT User Manuals and Tables of Services. The CAT manual made the conceptual model of the command language explicit. The user was told literally what the structure looked like while it was stressed that there were *groups* of services. On the other hand, the NoCAT manual was structured only implicitly (because of the preferred equivalence of codes, names, etc.). So, although there was a structure, on the surface it appeared that there was just a flat set of services with no salient interrelations.

5.2.2 Table of Services

As a consequence the Table of Services contained just one long list of services an example of which is shown in table 3. Note that, in order to avoid a syntax that was too explicit, it was necessary to portray activation and deactivation as "different services". In other words, every possible command string had a unique description that was associated with it.

5.2.3 User Manual

The NoCAT User Manual contained an introduction to the supplementary services and some information about the user interface. Besides that, it was explained that every service had a service name. However, there was no information on the syntax. Rather, it was stated that, for each service, after the ">"-sign there was a string that should be entered in order to (de-)activate a service. As mentioned before, the goal was to keep User Manuals and Tables of Services as minimal as possible, containing only essential information.

Clause 7 describes an experiment that was designed to test the effect on users' performance of these two different descriptions of the same command language. The next clause deals with the implementation of the command language and both design and implementation of an experimental PBI++ for accessing ISDN supplementary services.

Table 3: A typical section from the NoCAT Table of Services

Send on - "all calls" This service allows you to send on *everyone* who calls you up to another telephone automatically. For that purpose you must enter the number of that telephone (*the destination*), after activating the service. *How to use it?* Activation is possible only if you hear the dialling tone. >*21* **Stop sending on - "all calls"** This service allows you to stop sending on phone calls by means of the "Send on - all calls". *How to use it?* Deactivation is possible only if you hear the dialling tone. >#21#
Send on - " when busy" This service allows you to send on *everyone* who calls you up to another telephone automatically. For that purpose you must enter the number of that telephone (*the destination*), after activating the service. Phone calls are sent on ONLY if your telephone is BUSY. *How to use it?* Activation is possible only if you hear the dialling tone. >*22* **Stop sending on - "when busy"** This service allows you to stop sending on phone calls by means of the "Send on - when busy". *How to use it?* Deactivation is possible only if you hear the dialling tone. >#22#

6 An enhanced minimal phone based interface for ISDN supplementary services

In order to have a vehicle on which to implement the command language described in the previous clause, it was necessary to design an experimental phone based interface. As it had to be a PBI++ and no telephone hardware was available that allowed for rapid prototyping and wide variations in the initial design, it was decided to simulate the PBI++ by means of a touch screen. Subclause 6.1 lists a number of design considerations with respect to human factors aspects of the phone based interface, independent of the command language. Next, the implementation of the simulated PBI++ is described (subclause 6.2). Subclause 6.3 deals with a number of unavoidable differences between the CAT and the NoCAT interface.

6.1 Design

It is stressed here that design of a user interface was not the goal of this undertaking. Moreover, the reader should bear in mind that it was not a major goal to investigate the usability of these specific supplementary services. Rather, the research questions concern users' interaction with *supplementary-service-like* features. It was decided that ISDN supplementary services would serve as a vehicle for the present experiment at NL-PTT Research. It was merely necessary to develop some kind of interface which was suitable to study users' interaction with supplementary services. Therefore, ETSI stage 1 (service descriptions) and stage 2 (functional capabilities and information flows) of a core set of supplementary services (CLIR, COLP, CFU, CFB, AOC-S, AOC-E, HOLD, 3PTY and CBACK NR) were taken as basic descriptions for the design of a number of experimental user procedures. The ETSI documents contain device-independent descriptions of the supplementary services. The designer has to translate the descriptions into a design implementation.

Aside from a few exceptions (e.g. HOLD), user-system interaction always proceeded in the same way. Therefore, it was decided that there should be a uniform way to (de-)activate the supplementary services. The following design considerations have lead to the final prototype.

6.1.1 Visually displayed menus: parallel presentation of information is impossible

The definition of a PBI++ allows for a small display. It was decided that the size of the display should be 2 rows by 20 characters. However, this put a severe limitation to the type of information that could be displayed. For, instance it was nearly impossible to provide a menu with more than 2 items (e.g. service names). A solution might be to construct a "serial" menu, allowing the user to skip and scan items one at a time. However, in case of an appreciable number of items the lengthy summing up of, for instance, service names, would slow down the interaction too much. It is just the (visually) *parallel* structure of menus that make them so appealing. People are outstanding processors of information which is visually displayed in parallel. It was decided that the screen would be used for feedback only.

6.1.2 Auditory Instructions

If the user is not prompted by the screen, then a logical next step is to choose for auditory prompts. Auditory prompts are preferred above visual ones anyway, because the latter require the (ignorant) user to have his attention directed to the display at the moment of prompting and the former do not. So, for each step in the (de-)activation procedure the user was prompted by a female voice to press a button. Paradoxically, in a number of system states the auditory prompt consisted of a two choice menu. If the serial presentation of menu items is considered not to be appropriate, then voice menus should be avoided as well. Nevertheless, a two choice menu was regarded acceptable. In fact, there were only two menu options available: to progress or to take the whole procedure one step back. Besides, just as in HCI, there is a trade off between speed of the interaction and the amount of a priori knowledge the user has to have. The choice for a simple menu structure relieved the user from remembering "what to type next" and "how to go back". Still a smooth interaction was possible, partly due to the use of shortcuts as well (figure 7).

6.1.3 Recognition and aided recall vs. plain recall

It was stated in clause 4 that an important shortcoming of current phone based interfaces to supplementary service-like features is that they force the user to recall the commands. In order to (de-)activate a service, the user must establish two links: first, between the goal to be accomplished (establishing a certain state in the system) and the specific command or commands that accomplish that goal, and second, between the command (e.g. forward all incoming calls) and the key presses that initiate that command (Egly et al., 1985 [6]). In current phone based interfaces the relation between the task goal and the command that is entered seems to be entirely arbitrary to the user. This makes recall very difficult.

The use - next to codes - of service *names* which refer to, for example, the task goal that is accomplished by (de-)activating the service is a first step to a solution. In order to enter the right command the user should not remember an abstract code, but a meaningful name which corresponds to his task goal (e.g. an intended system state). However, it is only possible to enter *, # and digits. Therefore, the screen is used to display the service name as a form of feedback. This way the user can type a service code and see if the name that appears corresponds to his task goal or his remembrance of a previously encountered service name. The difference between recall and recognition or "aided recall" is illustrated by figure 5. Note that the lower right relation is realized in the user interface. This leaves only the upper right (semantic) relation to be maintained by the user.

As discussed in the previous clause, the command language does in fact have a menu-like structure. So, if the user succeeds in incorporating the intended conceptual model, one could argue that there is a "user-initiative menu-based" interface. No menu structure is presented in the interface, although a structure is there, but the user *is* prompted auditorily to choose an (invisible) "menu option" (service code). In other words, if he has succeeded in incorporating the conceptual model of the language, his mental presentation will "look like" some kind of two layer menu structure (layer 1 = groups; layer 2 = services). The service name is displayed as a form of feedback on the choice from the "mentally represented menu".

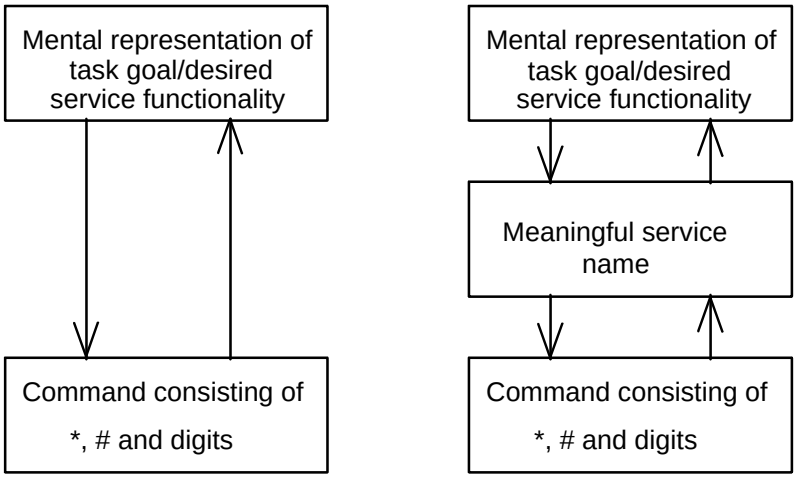

Figure 5: The difference between plain recall of a command string in current phone based interfaces (left) and recognition or "aided" recall in the newly designed PBI++ (right)

6.1.4 Use of shortcuts

A final yet important design consideration was the use of shortcuts. Users were expected to gradually acquire some level of expertise. As a consequence (parts of) frequently used service commands might become "chunks" of information that are typed at once. At all times it should be possible for the user to interrupt the display of auditory as well as visual information and proceed. In other words, novice and intermittent knowledgeable user were aided by the service names, while expert users were expected to skip the lexical part of the association eventually and form direct relations between service functionalities and commands. So, in the initial stage of user learning, the phone based interface should serve as a kind of menu-prompting interface, while - later on - this same interface would allow the typing of entire command strings in the same manner as current phone based interfaces do.

6.1.5 The final prototype

Figure 7 shows a typical user procedure of the prototype that was used in the experimental CAT condition, figure 8 shows the same procedure for the NoCAT condition. The following description is about figure 7. User actions result in the system status shown in the other two columns. Note that the two procedures in figures 7 and 8 differ with regard to the use of the group digit and group name.

As the user has decided to activate CFU (1), he first searches for the right category. He tries *3* for a group number (2). However this is not the group he is looking for. So, he decides to take the procedure one step back (3). Next, he types *2* and concludes that this must be the one (4). Then he tries *2* for a service number (5). Yet, *Send on - "when busy"* (CFB) is not the preferred service, so he goes back (6) and tries *1* (7). This service does match his goal. Therefore he decides to activate it by pressing * (8). He enters the phone number (9). Finally he presses * to conclude the activation procedure (10). It is important to note that at all the times system instruction and feedback could be interrupted by just executing the next action in the procedure. So, for example, it is possible to press *3#22#1*312781* (or part of it) and next wait and see what the result is.

As a concluding remark it should be pointed out, that the experimental user interface only uses 6-digit phone numbers. As a consequence it is possible to tell the user to *press **(10). A "real life" phone based interface should instruct the user to *type the preferred phone number, followed by a **at (9), because of the varying length of phone numbers. The reason for not employing this way of user instruction was, that it was feared that it would place too heavy demands on STM, especially for elderly novices. STM aspects of phone based interfaces should be a separate (important!) field of study.

6.2 Simulation

The experimental user interface and command language were simulated by means of a touch screen. A HyperCard® stack was created on an Apple® Macintosh® computer which displayed the image (figure 6) and (sampled) sounds of a telephone. The telephone could be operated by touching the screen at the site of the displayed buttons. Operation of the telephone proceeded in exactly the same way as a "normal" telephone (except for the supplementary services, of course). If it was on-hook, the only possible action was to lift the receiver (by touching it!). As soon as the receiver was lifted, the user heard the dialling tone. He could then start dialling a phone number, or enter the "supplementary services mode" (figures 7 and 8).

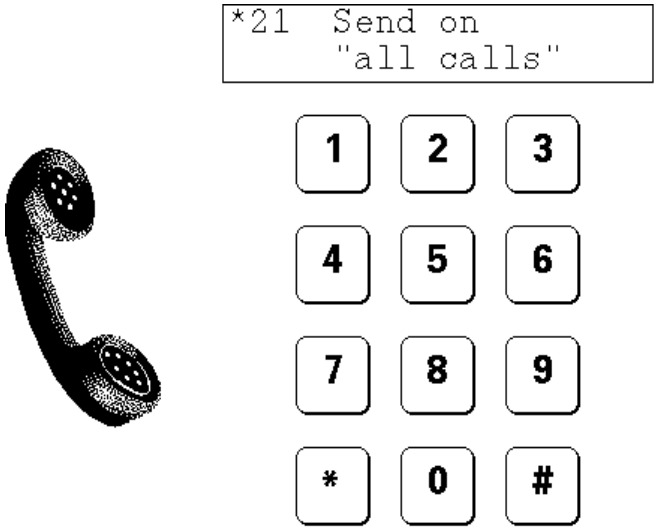

Figure 6: The image that was displayed on the touch-screen

Table 4: A typical user procedure in the CAT condition (activate CFU)

	User	System - instruction (voice)	System - feedback (LCD)
1	Lifts receiver Presses *	"To proceed, type the number of the preferred group" "To leave the supplementary services, press #"	* Activate service
2	Types 3	"To proceed, type the number of the preferred service" "To choose another group, press #"	*3 More than two
3	Presses #	"To proceed, type the number of the preferred group" "To leave the supplementary services, press #"	* Activate service
4	Types 2	"To proceed, type the number of the preferred service" "To choose another group, press #"	*2 Send on
5	Types 2	"To activate this service, press *" "To choose another service, press #"	*22 Send on "when busy"
6	Presses #	"To proceed, type the number of the preferred service" "To choose another group, press #"	*2 Send on
7	Types 1	"To activate this service, press *" "To choose another service, press #"	*21 Send on "all calls"
8	Presses *	"Please type the preferred phone-number"	*21* "all calls"
9	Types 321781	"Press * now, please"	*21*321781 "all calls"
10	Presses *	"The service has been activated" .....Dialling Tone.....	*21*321781* "all calls"

Table 5: A typical user procedure in the NoCAT condition (activate CFU)

	User	System - instruction (voice)	System - feedback (LCD)
1	Lifts receiver Presses *	"To proceed, type the number of the preferred service" "To leave the supplementary services, press #"	* Activate service
2	Types 35	"To activate this service, press *" "To choose another service, press #"	*35 More than two "Conference"
3	Presses #	"To proceed, type the number of the preferred service" "To leave the supplementary services, press #"	* Activate service
4	Types 22	"To activate this service, press *" "To choose another service, press #"	*22 Send on "when busy"
5	Presses #	"To proceed, type the number of the preferred service" "To leave the supplementary services, press #"	* Activate service
6	Types 21	"To activate this service, press *" "To choose another service, press #"	*21 Send on "all calls"
7	Presses *	"Please type the preferred phone-number"	*21* "all calls"
8	Types 321781	"Press * now, please"	*21*321781 "all calls"
9	Presses *	"The service has been activated" .....Dialling Tone.....	*21*321781* "all calls"

6.3 Differences between the CAT and the NoCAT User Interface

As the CAT manual and user-interface capitalized on the group-structure, while the NoCAT manual and user interface should have been flat, a difference between the CAT and the NoCAT user interfaces was inevitable. In the NoCAT user interface, "group terminology" should have been avoided. In fact the difference concerned only one stage in the user procedure. As a CAT user entered the group and service numbers one by one, NoCAT users were prompted to enter the "number of the preferred service" at once. In the NoCAT interface, the two digits together denoted the service number. In other words, NoCAT users skipped the step of entering a group digit. Auditory instructions as well as feedback and "soft-key-functionality (effect of * and #) were tuned to this slightly altered procedure.

7 Method

After discussing the operation and implementation of one of the main independent variables in the two preceding clauses, the next logical step is to design an experiment in order to test the hypothesis and answer the research questions which were formulated on the basis of the fourth clause.

7.1 Experimental Task

Subjects in the experiment had to solve a number of problems, using the simulated supplementary services. Each problem was embedded in a *scenario*. Scenarios started with a description of a situation. In this situation an artificial goal or intention was set for the subject. The task was to activate a supplementary service to establish the goal state. So, subjects had to match functions with the goal state, using the User Manual and Table of Services. Scenarios were constructed very carefully in order to make sure that the situations were easy to imagine for everyone. Subjects made one attempt to solve the problem. They had to do the scenarios in a fixed order and they could not return to a previous scenario. Figure 9 shows a typical example of a scenario. Note that subjects had to find out themselves which service to (de-)activate, using a manual. Therefore, part of the task consisted of problem solving, i.e. finding a match between the goal and the appropriate service functionality. Note that subjects were given more than one phone number. This was done in order to avoid the possibility that users inferred which service to use by looking at whose phone number was presented.

Table 6: A (CFU) scenario that was used in the experiment

Description of the situation: Imagine you are going on a holiday. You are expecting an important phone call. A friend has offered to handle all incoming calls. However he cannot sit and wait at your telephone all day. It would be most practical if he could stay at home and receive all your incoming calls at his telephone. **Instruction:** Make use of a supplementary service, in order to make it possible for your friend to stay at home and answer your calls. Your phone number at home is **139753** The phone number of your friend is **638752**

7.2 Experimental Design and Independent Variables

A 2 by 2 factorial repeated measurements (between subjects) design was used to study the aforementioned expectations and questions (clause 4). The factors were type of transmitted conceptual model (i.e. type of manual, either CAT or NoCAT) and age group (young or elderly). A between subjects design was used. Therefore, one half of the subjects received the CAT manual and the other half received the NoCAT manual.

7.2.1 Supplementary services and scenarios

Effects of learning were examined within 4 supplementary services (CLIR, CFU, AOC-S, 3PTY). Each service was to be used in 4 functionally equivalent scenarios, which differed on the surface only. The scenario *type* was determined by the supplementary service that should be activated (e.g. a "CFU scenario"). In order to prevent subjects from discovering any pattern at an early stage of the experiment, for each type of scenario there were 2 dummy scenarios. In a dummy scenario a supplementary service was needed which resembled (same category) one of the four services in question. Finally, it was decided that before the first 3PTY scenario, there had to be a scenario in which the subject used HOLD separately. It served to let the subject familiarize with HOLD and avoid the need to use two new services in the first 3PTY scenario. This made up a total of 25, divided into a first block of 12 and a second block of 13 scenarios (table 7).

Table 7: The order of scenarios in the experiment

1. CLIR	6. COLP	11. CFB	16. AOC-S	21. 3PTY
2. CFU	7. CFU	12. AOC-E	17. 3PTY	22. CBACK NR
3. AOC-S	8. AOC-S	13. 3PTY	18. CFB	23. CFU
4. HOLD	9. 3PTY	14. CLIR	19. COLP	24. AOC-S
5. CBACK NR	10. CLIR	15. CFU	20. AOC-E	25. CLIR

CLIR is a service that allows the user to hide his identity. CFU transfers all incoming calls automatically to the indicated number. AOC-S can be used to request information on the costs of a call to a certain number, before the call is actually made. HOLD puts a call on hold ("waiting room"). CBACK NR was a (not really existing) service that could be used in case the called party was not at home. It would try to contact the called party every 10 minutes and upon succeeding it would call back the user. COLP allows the user to view the number of the person he finally reaches (this might differ from the number he has dialled, e.g. in case his call has been forwarded). 3PTY means three party calling. CFB can be used to direct incoming calls to another number, but only if the user is busy. Finally, AOC-E is used to request information on costs of a call that has been completed.

7.2.2 Type of conceptual model - CAT vs. NoCAT

The first factor in the experiment was the type of conceptual model which is communicated to the user through the user manual. The two levels of this factor differ in two ways:

- corresponding User Manual and conceptual model. This was either the CAT manual which capitalized on an explicit structure of the command language or the NoCAT manual which left the structure implicit (see clause 5);

- implementation of the user-interface. As a result of the fact that in the NoCAT condition the surface structure had to be that of a flat model (no groups), there were some slight differences between the user interfaces in both conditions (see clause 6).

7.2.3 Age

Secondly, two different age-groups were formed. In defining their boundaries, care was taken that the younger subjects had experience with all kinds of information technology (i.e. they should be young enough, 18-30). The older age group should consist of elderly whose age was such that they had not been using information technology intensively during the adolescent and professional periods of their lives (55-65).

7.2.4 Subjects

Younger subjects were recruited by an employment agency. Elderly subjects were selected from a university pool. A number of the elderly subjects had participated in a "Teleshopping" experiment a few months before. In assigning the subjects to conditions, care was taken that education and sex were equally spread, especially in the younger age group. With regard to the elderly, education was considered less important than (former) occupation and experience with information technology, although an attempt was made to match younger and elderly groups on education as well. The younger subjects received an hourly wage. The elderly were given a fixed fee, independent of the duration of the experiment.

7.3 Data registration

7.3.1 Performance measures

7.3.1.1 Video-registration

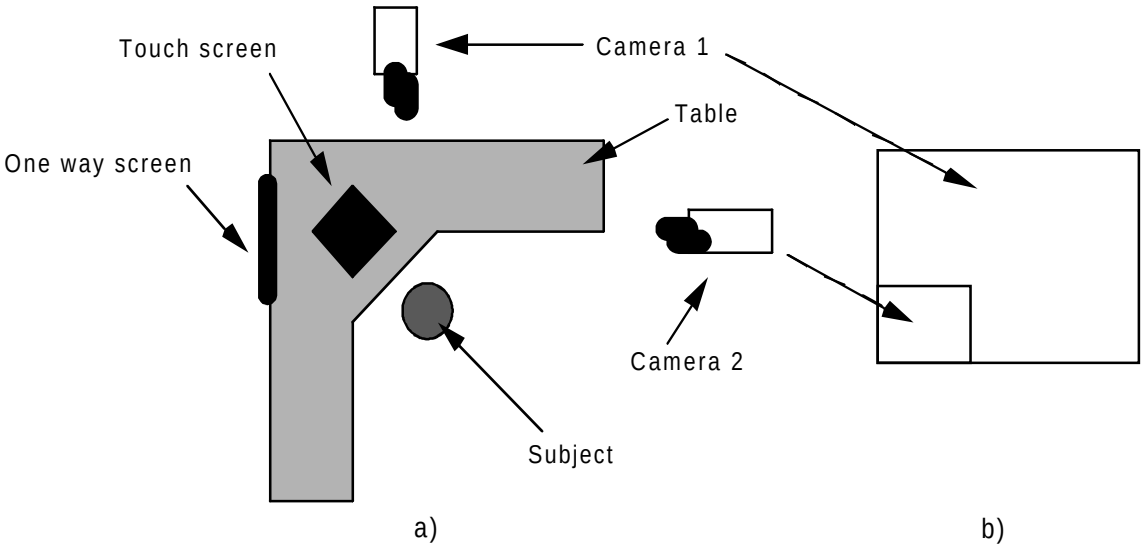

Figure 7: a) The video-equipment set-up; b) The combination of the two views in one single video window

Those parts of the experiment during which subjects were actually working with the supplementary services were recorded on video. Figure 7a) shows the video-equipment set-up. Camera 1 was used to record the subject's manual consultations and facial expression. Camera 2 registered the subject's keying behaviour. Both views were put together in one single video-window (figure 7b).

7.3.1.2 Observations

The time subjects took to consult the manual was recorded with a chronometer by on-line observation, either through the one way screen (figure 7a) or by observing the on line video monitor, dependent on which allowed the best view of the subject's gaze. Duration of manual consult were recorded on an observation form. Any exceptional behaviour was recorded on the same form.

7.3.1.3 Logging of subject's actions

The HyperCard® implementation used on an Apple® Macintosh® made it possible to log keying actions. Table 8 contains a list of all possible events that were recorded.

Table 8: All possible events that were recorded

digit	subject pressed a digit button
*	subject pressed *
#	subject pressed #
START 1 2	start of scenario number 1 scenario ID 2
UP	subject lifts receiver
DOWN	subject puts receiver down
BACK	back to previous scenario, in case a subject accidentally pressed the scenario button twice
RESTART	in case the program had terminated unexpectedly

7.3.1.4 "Model test"

As mentioned before, users have trouble remembering service codes. To test subjects' recall of service codes, a memory test was designed. It consisted of 19 questions in the NoCAT condition and 26 questions in the CAT condition (additional questions on group numbers). Subjects were tested for retention of service codes, (de-)activation procedures and formation of an explicit mental representation of the underlying conceptual model.

7.3.1.5 "Repeated model test"

Subjects' recall over a longer period of the time was investigated as well. For that purpose they received exactly the same test, two weeks after the initial experiment.

7.3.2 Personal background data

7.3.2.1 Structured interview

In order to form a notion of the subject's social, educational and professional background and previous experience with information technology, telephony and supplementary services, a structured interview was held. These data might serve to explain any variability in behaviour.

7.3.2.2 Standard progressive matrices

A non-verbal reasoning test was used for all subjects. There were two reasons for this. Firstly, it was necessary to test whether subjects in one condition had an equal level of abstract intelligence as subjects in the other condition. Secondly, it is interesting to see whether there is a relation between performance, level of education and ability to solve abstract problems. The Standard Progressive Matrices (SPM) section of the Raven test was chosen for this purpose. The SPM was constructed to measure a person's deductive ability. The essential feature of deductive ability is the ability to generate new, largely non-verbal, concepts which make it possible to think clearly (Raven et al., 1992 [24]). An important

consideration in choosing the SPM was, that strategies required to solve the problems in this test resembled those required for solving the problems in the experiment in some important ways. The SPM consists of five series of 12 problems. Each problem is completed by choosing an abstract visual diagram-structure that fits within a series of others. In other words, to solve a SPM-problem a subject has to understand a complex and relatively abstract structure (corresponding to a "situation") and find an element (corresponding to a "supplementary service") that completes the structure. However, there is one important difference between these two tasks. The SPM is non-verbal, while the experimental task is highly verbal, because subjects have process a lot of written information (scenario descriptions and user manual). This allows to test whether subjects who perform badly in the experiment do so, because they do not read or process Dutch texts very well. In case of a high SPM score and low performance scores in the experiment, it may be concluded that this performance discrepancy is at least partly caused by the subject's ability to process written information.

The Standard Progressive Matrices part of the Raven test is used to measure just one aspect of intelligence or intellectual capacity. Usually it is employed in combination with the Mill Hill vocabulary scale. Although formally not entirely correct, in this report performance on the Standard Progressive Matrices will be referred to as *abstract intelligence* or *abstract ability*.

7.3.3 Subjective measures

7.3.3.1 Subjective mental effort

An additional performance measure consisted of subjective evaluations of the mental effort invested in carrying out the experimental task. A learning effect might be accompanied by a decrease in mental effort. For this reason, subjects completed the Effort-scale (Zijlstra & Meyman, 1989 [35]). It consists of a scale which runs from 0 to 170. Along the scale different indications are marked ranging from "totally no effort at all" to "very much effort". Subjects mark the statement they consider most applicable.

7.3.3.2 Acceptance questionnaire

Usability is not only a matter of objective performance, or subjective mental effort. Although these two bear a close relation to it, user *acceptance* is a third major topic of interest. For marketing purposes it is useful to know what projected users think of a product. Usable products are not necessarily marketable and vice versa. In other words, a user may be able to use it, but still not find the interaction pleasurable. Besides that, a user friendly system is not necessarily *useful.* To measure user acceptance, a questionnaire concerning the aforementioned issues was constructed.

7.4 Definition of dependent variables

On the basis of the data that were recorded a number of dependent variables were defined, which were used to statistically test the hypothesis formulated in the next subclause.

7.4.1 Duration of scenario

The start of a scenario was defined by the moment the subject put down the receiver for the last time in the previous scenario. The end of a scenario was defined either by the moment of completion of the appropriate commands or by the moment the subject put down the receiver for the last time. The last criterion was used in case the subject did not manage to activate the right service(s). Duration of a scenario is the time between the defined start and end of it. For scenarios 1 and 13 there are no previous scenarios. Therefore, start of these is defined by the moment a subject starts reading the description of the scenario. Total duration of all scenarios was also taken as a performance measure. It was calculated by summing the duration of all 25 scenarios.

7.4.2 Duration of manual consultation(s)

If there were more than one manual consultation, all durations of single consultations were aggregated into one duration for the entire scenario. The start of a manual consultation was defined as the moment the subject looked into the User Manual or Table of Services. The end was defined as the moment he started looking at the touch screen again. Total duration of manual consultations over all scenarios was also taken as a performance measure. It was calculated by summing the duration of manual consultations over all 25 scenarios.

7.4.3 Errors

For each scenario, the subject had to perform a number of mental operations. In order to be able to act in the right way the subjects had to:

1) decide whether to operate any supplementary service at all;

2) in case of operating a service, decide which service and remember the code;

3) decide (remember) on what moment it should be (de-)activated (before, during or after a call had been made);

4) remember the syntax for (de-)activating the service.

On the basis of this rudimentary "task-analysis" a taxonomy of errors was constructed which is presented in table 9. Error types were indicated by a 4-digit score, one digit for each decision the subject had to make. For each scenario there could be one or more error codes. For instance, for three-party scenarios there were two error codes: one for the activation of HOLD and one for the activation for 3PTY. Analyses were performed on number of errors and type of error. Total number of errors was also taken as a performance measure. It was calculated by counting the number of wrong solutions over all 25 scenarios.

Table 9: The taxonomy of error types that was used to construct error scores

1st digit	2nd digit	3rd digit	4th digit
Service activated?	Correct service code?	Correct moment?	Correct "syntax"?
0 = no	0 = no	0 = no	0 = totally wrong
1 = yes	1 = yes	1 = yes	1 = yes
	2 = correct category		2 = activated instead of deactivated or vv.
			3 = 1st * or # correct
			4 = wrong phone number
			5 = others, not serious

7.4.4 Level of "abstract intelligence"

The raw SPM scores (total number correct out of 60 problems) were taken as a measure of subjects' intellectual capabilities. It was noticed that SPM scores decrease slightly with age, however it was just this aspect of ageing that was of interest.

7.4.5 Model test scores

The model test answers were rated using the same way of scoring errors as for performance. So, for each answer in the model test there would be an n-digit score, n depending on the type of question. For instance, there were questions on service codes. In case of such a question the error score would consist of only one digit, indicating whether the right service code was chosen (score = 1), only the group number was correct (score = 2), or the answer was entirely wrong (score = 0). Other questions dealt with entire activation procedures (e.g. establish a three party call). For those questions there were 4-digit scores. The model test score was calculated by counting the number of totally correct answers. It should be noted that the CAT model tests contained some additional questions on grouping of services, and group numbers. Therefore, model test scores were constructed only on the basis of the questions that both conditions shared.

7.4.6 Treatment of missing data

Missing data are the result of, for example, subjects skipping or not completing a scenario, interruption of the scenario in case of serious problems that endanger continuation of the experiment. With regard to duration of scenarios and duration of manual consultations, the missing values were replaced using the SPSS command "RMV *newvar* = TREND(*oldvar*)" which computes the new value on the basis of a regression analysis of trends in the remaining non-missing data.

7.5 Procedure

The procedure consisted of a number of separate stages. Each stage started with an instruction which was read out aloud to the subject. Except for filling in the second Model test, the procedure took place in the Usability Lab at NL-PTT Research. Table 10 shows the main stages in the procedure.

Table 10: The main stages in the experimental procedure

-	General introduction
-	Structured interview
-	Introduction of touch-screen and practice period
-	User Manual and Table of Services
-	First part of the experimental task
-	Effort-scale 1
-	Second part of the experimental task
-	Effort-scale 2
-	Acceptance questionnaire
-	Standard Progressive Matrices
-	Model Test
-	Repeated Model Test

The procedure started with a general introduction. The purpose of the experiment (usability study) was explained. An overview of the procedure was given and at the end the subject was asked to fill in a permission form concerning the videorecording. Next, the experimenter and the subject completed the first questionnaire together. Upon completion of this structured interview the subject was guided to the touch-screen on the other side of the room. It was explained that the "printed" telephone could be operated by touching the screen. Thereafter, the subject got a number of training exercises to perform. He was left alone by the experimenter.

It was noticed that subjects had to process a lot of information at once. To prevent information overload at an early stage in the experimental procedure, subjects were instructed to read the User Manual first. However, the Table of Services was available to them all the time (partly because the introductory part of the User Manual referred to it and partly to give them a chance to see how the services looked. They might start to read the Table whenever they felt they were ready to). The subject was left alone to allow him to read the User Manual quietly. As soon as he had finished reading, the experimenter entered the room again and instructed the subject to skip and scan the Table of Services. The last instruction before the first 12 scenarios consisted of an explanation of "how to do" the scenarios and how the subject could contact the experimenter in case of a question or other problems. The subject then received the first 12 scenarios, preceded by a short introduction. Finally, the subject was urged to use the User Manual and Table of Services as little as possible. The experimenter left the room and the subject started working.

7.5.1 Time limit

With regard to the duration of the experiment it was necessary to define a time limit. The total duration of completing the first 12 scenarios was taken as an index of subjects' speed of performance. If the subject did not succeed in completing the first 12 scenarios within the hour, the second part of the experiment (13 scenarios) was skipped and the procedure was continued by filling in the acceptance questionnaire.

After 12 scenarios or when the time limit had been reached, the experimenter entered the room again. The subject marked the Effort scale immediately. If he wanted to, a short break was possible. After filling in the Effort-scale the subject was given the next part of the scenarios. He was asked to proceed in the same way as during the first part of experiment.

Upon finishing the second part of the experiment, the subject marked the Effort-scale once again. Next, he was guided to a quiet room for the last part of the procedure. There the subject was given the acceptance questionnaire. Again the subject was left alone, but he could always contact the experimenter in the same ways as during the other parts of the experiment. when the subject had finished the acceptance questionnaire, the experimenter introduced the SPM and completed the standard instructional procedure (Raven et al., 1992 [24]). Finally, the first Model test was introduced. Once again it was stressed that it was not the subject who was being tested, but this time the clarity of the equipment in facilitating recall of the operational procedures. The Model test was preceded by a short introduction as well. It is important to note that there was about one hour between completion of the last scenario and the

Model test. Subjects were unaware of the fact that they would be given a test. Besides, subjects were too busy completing the acceptance questionnaire and the SPM to be able to actively remember the service codes. After the subject had completed this final questionnaire, he was invited to participate in a follow-up survey. All subjects agreed and filled in their name and address. Exactly two weeks after participating in the experiment, subjects received the same Model test once more by mail. Returning the questionnaire was rewarded by sending them the results of the SPM, if the subject so wished.

7.6 Pilot study

The above-mentioned experimental task and procedure are based on a pilot study which was carried out with 7 unpaid students. They participated voluntarily and completed the procedure as it was developed initially. It emerged that, although 6 of them were able to solve the problems quite easily, there were too many scenarios (38). The one subject (CAT condition) who was not able to solve the problems was extremely confused and unable to form a model. This was partly due to the fact that he thought he had to enter a 3 digit code (e.g. group number *2*, service number *21*, resulting in service code *2 21* instead of *2 1*). Earlier versions of the CAT Table of Services contained the service code ("absolute address") of each service instead of the service number ("rank within the group"). It was decided that each service would be given a 1 digit number in the manual, so that subjects had to construct the service code by combining the group number and the service number. The number of scenarios was lessened to the current 25. For the rest, some small - yet important - details in the texts of the User Manuals, Tables of Services, the user interface itself and the scenarios were changed.

7.7 Hypothesis

The hypothesis and research questions formulated in clause 4 can now be reworded in terms of their operations, i.e. the experimental variables.

H1 *Duration of manual consultations and duration of scenarios will decrease with time/scenario.*

H2 *As an accompanying effect, subjective mental effort ratings for the first part of the experiment will be less than for the second part.*

H3 *Total duration of task completion, total duration of manual consultations, and total number of errors will be less in the CAT condition than in the NoCAT condition.*

H4 *The decrease proposed in H1 and H2 will be stronger in the CAT condition than in the NoCAT condition.*

H5 *Total duration of manual consultations, total duration of all scenarios together and total number of errors will be less in the younger condition than in the elderly condition.*

H6 *The decrease proposed in H1 and H2 will be stronger in the younger condition than in the elderly condition.*

H7 *Elderly subjects will give higher subjective mental effort ratings to both the first and second part of the experiment.*

H8 *There will be a negative relation between SPM score and total duration of manual consultations, SPM score and total duration of all scenarios together and finally SPM score and total number of errors.*

H9 *Performance on model test I will be better in the CAT condition, i.e. subjects will remember more service codes and (de-)activation procedures in a correct way. Performance differences between CAT and NoCAT will be larger on model test II than on model test I. In other words, one hour after completing the last scenario, CAT-subjects will remember more than NoCAT subjects, but two weeks after the experiment this difference will have grown.*

7.8 Statistical Analyses

In order to test the aforementioned hypothesis, statistical tests were performed, using a common PC-based statistical package (SPSS), results of which are presented in the following clause. With regard to the hypothesis which state the direction of effects, a one-tailed less than 5 % type I error probability was considered to constitute a statistically significant effect. The exact tests that were used are described in the next clause.

8 Results

This clause deals with the main results of the experiment. First, a number of global observations are presented. Subclause 8.1 deals with the (necessarily descriptive) results concerning the elderly. The last subclause (8.2) starts with descriptive statistics as well (younger subjects). Finally, the results of statistical tests on effects of presenting different types of conceptual model (the second experimental factor) are presented, as well as effects of practice.

Upon arrival at the NL-PTT Research Usability Lab, it appeared that one 37 year old and one 52 year old female had been selected by the employment agency. Because they did not belong to either age group, it was decided to remove their data from the present analyses. The remaining 32 subjects participated in the experiment. Mean age in the elderly subgroup was 58,0 (sd = 2,71), mean age in the younger subgroup 23,1 (sd = 3,4). The distribution of subjects and sexes over conditions is shown in table 11.

Table 11: The assignment of subjects to conditions

	CAT		**NoCAT**		
	male	**female**	**male**	**female**	
younger	5	4	5	5	19
elderly	5	2	5	1	13
	16		16		32

There were 13 subjects who did not complete the entire experiment because they exceeded the time limit. Table 10 shows the number of subjects who did complete 25 scenarios and their distribution over conditions.

Table 12: Number of subjects who completed all 25 scenarios (total number of subjects in parentheses)

	CAT	**NoCAT**	
younger	8 (9)	8 (10)	16 (19)
elderly	2 (7)	1 (6)	3 (13)
	10 (16)	9 (16)	19 (32)

It emerges from table 12 that there were too few elderly left to perform any statistical testing with regard to the hypothesis involving the elderly. Therefore, in the next subclause a number of descriptive statistics are presented, along with an informal test on mean duration of scenarios. Subclause 8.2 deals with statistical testing of hypothesis regarding conceptual models and effects of practice.

8.1 Descriptive statistics with regard to the elderly

It should be noted that completing 25 scenarios was an exception in the elderly age group. Only 3 males out of 13 elderly completed the entire experiment. Their performance scores are listed in table 13. Total duration stands for total duration of completing the task. The time limit was 60 minutes for the first 12 scenarios. The last column shows mean scores for 16 subjects in the younger age group who completed the entire experiment.

Table 13: Performance scores of the 3 male elderly subjects who did complete 25 scenarios (standard deviations in brackets)

Subject No	22	28	32	Mean younger
Condition	NoCAT	CAT	CAT	**-**
Total duration	72 m	63 m	72 m	69 m (16 m)
Total duration of manual consultations	14 m	29 m	19 m	22 m (8 m)
Total number of errors	2	3	2	4,31 (3,93)
Model test I	12	7	12	10,13 (3,58)

From the data presented in table 13 it may be concluded that the performance of the 3 elderly who did complete 25 scenarios in fact resembles that of the younger subjects. It should be noted that two of them where highly educated (academic), had staff functions and worked with computers and PBXs on a daily basis. The third, who had worked as a laboratory assistant, rated himself as quite a good (computer-) chess player. However, for two of them the subjective invested effort was not within the standard deviation around the mean of the younger age group (table 14).

Table 14: Effort scores of the 3 male elderly subjects who did complete 25 scenarios.

Subject No	22	28	32	Mean younger
Condition	NoCAT	CAT	CAT	-
Effort 1	40	73	70	47,13 (18,99)
Effort 2	29	74	60	37,52 (16,72)
NOTE: The last column shows mean scores for 16 subjects in the younger age group who completed the entire experiment (standard deviations in brackets)				

For 2 elderly women the task was so difficult that they had to be guided by the experimenter in solving the problems. Their behaviour might be subject of in-depth analysis in a secondary data analysis. For the current experiment, however, they are excluded from analyses of performance. The mean number of scenarios completed by the remaining 8 subjects was 9,75 out of 25.

As a consequence of the results presented above, it was not possible to statistically test the hypothesis concerning the elderly. It was considered not useful to invite more than the 13 subjects that had already participated. Tests for any effects of transmitted conceptual model, learning, etc. were skipped. Moreover, it was simply impossible to compute the main performance measures. As an additional non-formal analysis for both age-groups the *mean duration per scenario* was computed over all scenarios that were completed over all subjects within each age-group. Figure 8 shows the results, which illustrate the fact that there was a significant difference ($p < 0,02$) between the groups, the younger subjects taking less time. The vertical axis represents time (minutes).

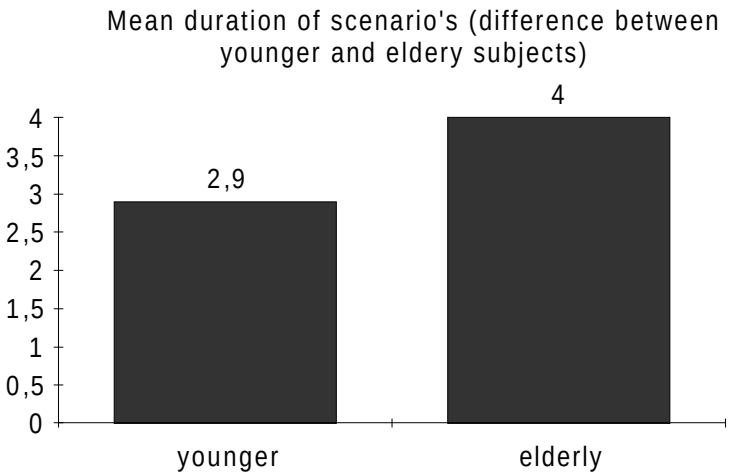

Figure 8: Differences between the younger and the elderly age-group in mean duration of performing a scenario, averaged over all scenarios that were completed

8.2 Effects of different types of transmitted conceptual model and effects of practice

There were 3 younger subjects who did not complete all scenarios. Their results are listed in table 15. Subject 07 solved several problems in a correct way, but took such a large amount of time that there was no time left to complete the second part of the experiment. He did in fact grasp the model of the command language, but did not understand the functionality of the services involved. Neither did he know when to activate a service (before, during or after a call). Subject 11 did operate the services correctly on a number of occasions, although the number of problems was even smaller than for subject 07. Finally, subject 20 seemed to lack motivation, partly due to a low success rate. For this subject, there seemed to be no intrinsic reward in solving the problems. Finally, subject 20 showed an inconsistent pattern of answers on the SPM. Note that the two available SPM scores are relatively low (48 is the boundary between 0,25 and 0,50 percentile). These 3 subjects were excluded from analyses of effects of conceptual model, because the period of time they worked with the services was not comparable to that of the other younger subjects.

Table 15: Data of younger subjects who did not complete all 25 scenarios

Subject No	**07**	**11**	**20**
Condition	CAT	NoCAT	NoCAT
Total duration (note)	69 m	58 m	53 m
Total number of scenarios	12	12	12
SPM	47	51	N/A.
NOTE: Total duration refers to the total duration it took to complete their scenarios (12)			

Finally, there were 16 younger and 3 older subjects whose data were entirely valid. It was decided not to join the elderly to the younger group, because there was quite a large difference in subjective effort which indicates that the three elderly people's and younger subjects' data were not entirely comparable. Although individual differences in subjective mental effort are often found to be large, the fact that the elderly were simply of a different subgroup was another concern. In other words, it is not known *exactly* what caused them to perform so well. The analyses described below employ data of the 16 younger subjects who did complete the entire experiment. The analyses are described for the main research questions and hypothesis formulated at the end of the previous clause. The number of cases is 16 (n_{cat} = 8; n_{nocat} = 8), unless stated differently.

8.2.1 Duration of manual consultations and duration of scenarios will decrease with time/scenario

In order to test the hypothesis that performance speed increases with practice, a Pearson *r* correlation coefficient was calculated between duration of scenarios and scenario number and duration of manual

consultations and scenario number. Table 16 shows the results. Duration of a scenario as well as duration of manual consultations decrease significantly ($p < 0,001$) with time/scenario number.

Table 16: Correlation between duration of scenario and scenario number and duration of manual consultations and scenario number

n = 400 (16x25)	duration of scenario	duration of manual consultations
scenario number	$r = -0,29$ ($p < 0,001$)	$r = -0,28$ ($p < 0,001$)

8.2.2 Subjective mental effort ratings for the first part of the experiment will be less than for the second part

The subjective mental effort as measured by the Effort scale should decrease as well. A MANOVA revealed a significant decrease ($p < 0,05$) on the within subject factor representing effort over both conditions (figure 9).

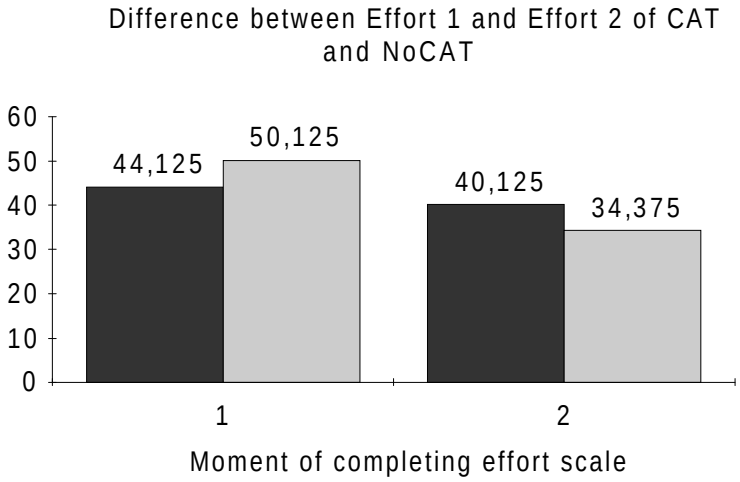

Figure 9: The difference between subjective mental effort after the first and second part of the experiment was significant ($p < 0,05$). CAT is displayed dark-grey. Subjective mental effort (maximum value 170) is displayed vertically

8.2.3 Total duration of task completion, total duration of manual consultations and total number of errors will be less in the CAT condition than in the NoCAT condition

An ANOVA was performed on total duration of task completion, total duration of manual consultations and total number of errors, with manual condition as a factor. There was no significant difference in mean total duration of completing all scenarios. The difference in mean total number of errors was also not significant. However, CAT subjects took significantly shorter to consult the manual ($p < 0,03$). The results are shown in figures 10, 11 and 12.

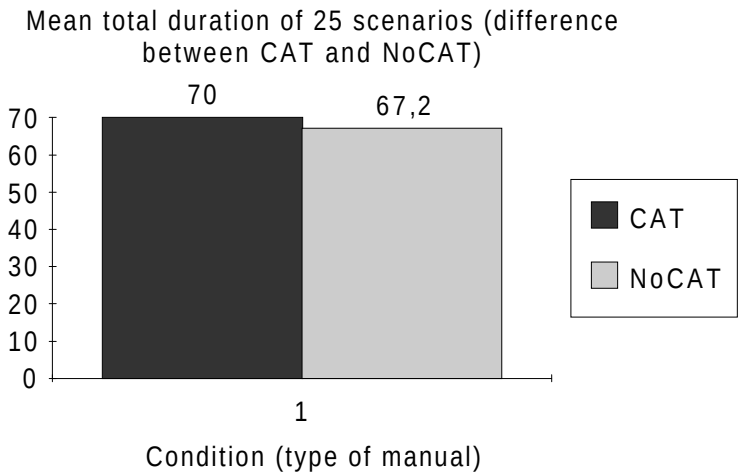

Figure 10: Differences between CAT and NoCAT with regard to mean total duration of task completion (n.s.). Time (in minutes) is displayed vertically

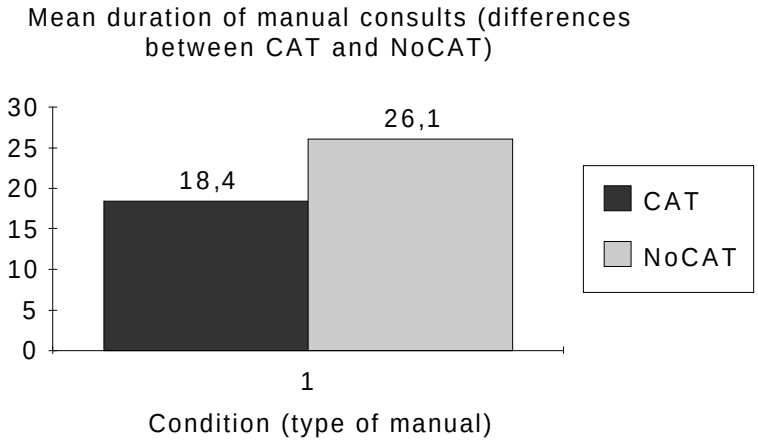

Figure 11: Differences between CAT and NoCAT with regard to mean total duration of manual consultations (p < 0,03). Time (in minutes) is displayed vertically

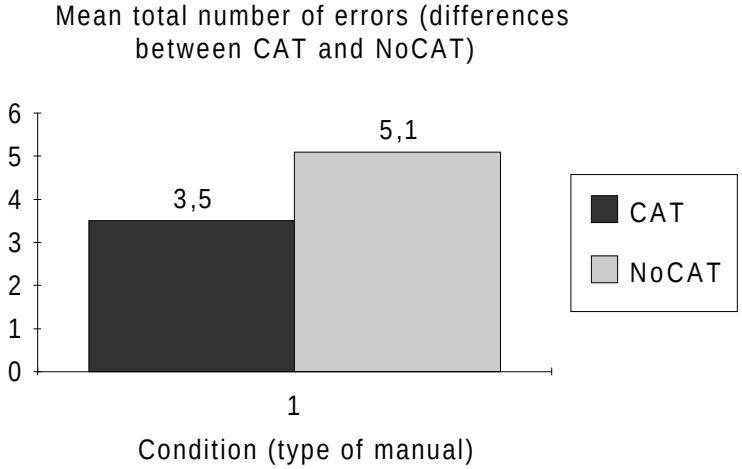

Figure 12: Differences between CAT and NoCAT with regard to mean total number of errors (n.s.). Number of errors is displayed vertically

8.2.4 The decrease in duration of scenarios and duration of manual consultations as well as the decrease in subjective mental effort will be stronger in the CAT condition than in the NoCAT condition

To test whether the learning effect was stronger in the CAT condition, four new variables were constructed. For the four services that returned four times, a score was computed for each occurrence of the services. So, a score was computed by summing the results of its first occurrences, another for its second, one for the third and one for the fourth. This was done for duration of task completion, duration of manual consultations and errors (table 17).

Table 17: The construction of 4 new variables for the three main performance measures duration of scenarios, duration of manual consultations and number of errors

D1 =	duration of 1st occurrence of CLIR + duration of 1st occurrence of AOC-S + duration of 1st occurrence of 3PTY + duration of 1st occurrence of CFU
D2 =	duration of 2nd occurrence of CLIR + duration of 2nd occurrence of AOC-S + etc.
D3 =	duration of 3rd occurrence of CLIR + etc.
D4 =	idem

In the same way four new variables were constructed for duration of manual consultations (M) and errors (E). Consequently there was a four level within subject factor for the learning effect. A MANOVA did not reveal any significant differences in performance between the CAT and NoCAT conditions on any of the dependent variables. There was also no difference between conditions with respect to the effect of practice. However, the performance changes aggregated over conditions found earlier were confirmed and this time a significant ($p < 0,03$) decline in number of errors was found too (table 18).

Table 18: Results of the MANOVA with regard to the 4-level within subject factor for practice/learning

	duration of scenarios	**duration of manual consultations**	**errors**
MANOVA	$F = 37.45$; $p < 0,001$	$F = 30.74$; $p < 0,001$	$F = 2.97$; $p < 0,03$

Figures 13, 14 and 15 show mean scores for both conditions on the four newly constructed variables for respectively duration of scenarios, duration of manual consultations and number of errors.

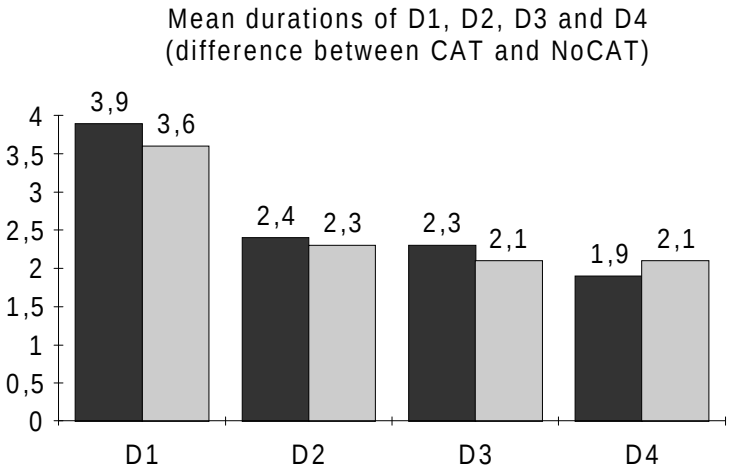

Figure 13: Mean duration of scenarios for CAT (dark-grey) and NoCAT on 1st, 2nd, 3rd and 4th occurrence of services. Time (in minutes) is displayed vertically

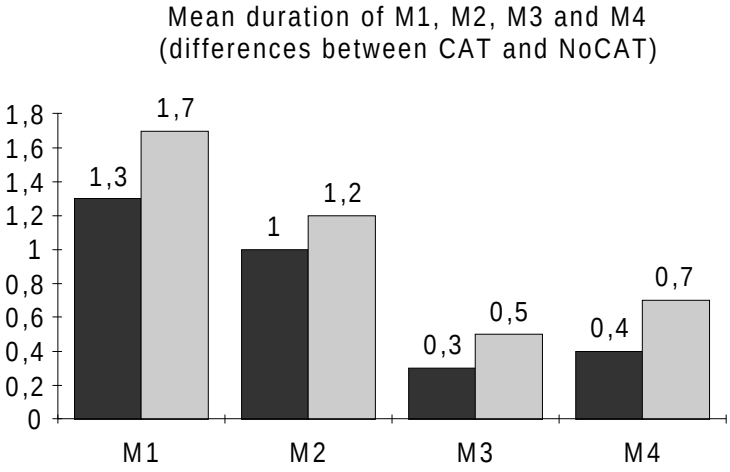

Figure 14: Mean duration of manual consultations for CAT (dark-grey) and NoCAT on 1st, 2nd, 3rd and 4th occurrence of services. Time (in minutes) is displayed vertically

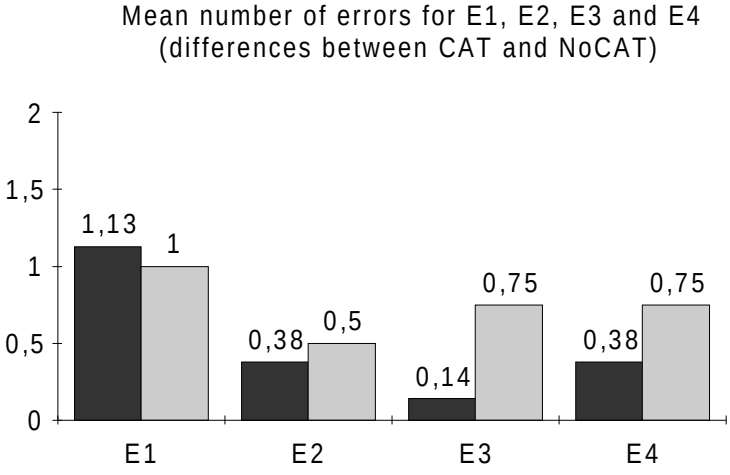

Figure 15: Mean number of errors for CAT (dark-grey) and NoCAT on 1st, 2nd, 3rd and 4th occurrence of services. Number of errors (maximum is 5) is displayed vertically

Data with regard to the hypothesis, that decreases in subjective mental effort differed in strength between both conditions, were presented above (figure 9). There were no differences between conditions.

8.2.5 There will be a negative relation between SPM score and total duration of manual consultations, SPM score and total duration of all scenarios together and finally SPM score and total number of errors

The relation between total duration of task completion and SPM score, total duration of manual consultations and SPM score and total number of errors and SPM score was tested by calculating a Pearson *r* correlation coefficient. The results are listed in table 19. Correlations between total duration and SPM score and total number of errors and SPM score were significant and relatively large. However, the correlation between total duration of manual consultations and SPM score was not.

Table 19: Correlation between total duration and SPM score, total duration of manual consultations and SPM score and total number of errors and SPM score, independent of condition

	total duration	total duration of manual consultations	total number of errors
SPM score	*r* = -0,52 (p < 0,02)	*r* = -0,19 (n.s.)	*r* = -0,55 (p < 0,02)

8.2.6 Performance on model test I will be better in the CAT condition; performance differences between CAT and NoCAT will be larger on model test II than on model test I

Differences between CAT and NoCAT subjects' scores on the first model test were tested by carrying out an ANOVA on model test scores with manual condition as a factor. CAT subjects remembered significantly (p < 0,05) more service commands in a correct way than NoCAT subjects. Figure 16 shows the mean scores.

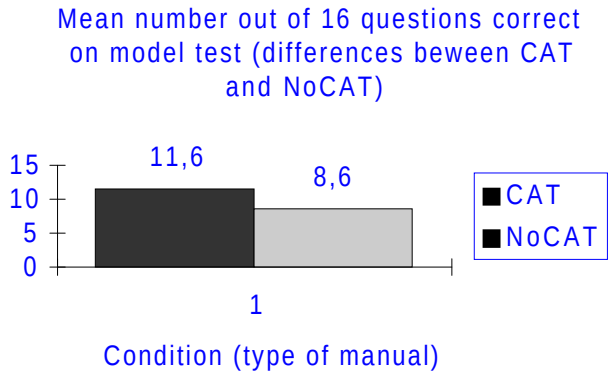

Figure 16: Differences between CAT and NoCAT on model test I (p < 0,05). Number of questions correct is displayed vertically

Calculation of model test II scores depended on the subjects' willingness to return the questionnaires. Table 20 shows the number of subjects who returned the questionnaire. As argued before, model test scores of subjects who did not complete enough scenarios were considered not valid.

Table 20: Number of subjects who returned the repeated memory test. Only subjects who completed 12 scenarios or more received the test (number in brackets)

CAT	NoCAT
7(8)	4(8)

In the CAT condition, nearly everyone returned the questionnaire. However, in the NoCAT condition only half of the subjects did, resulting in too small a number of subjects in each experimental condition. Therefore, it was not possible to test for any effect of manual type on recall of commands over a longer period of time. Table 21 shows the number of questions answered correctly on the first and second model test.

Table 21: Number of questions answered correctly on the first an second model test by younger subjects who returned the second model test (n = 11)

n = 11	model test I	model test II	difference
mean score	10,7 (3,8)	8,5 (3,4)	2,2 (2,6)

Subjects answer 2,2 more entirely correct questions on the first model test than on the second. In other words, decline in recall of service commands and procedures is 21 % over a two week period.

8.3 Types of error

Although no hypotheses were formulated in this respect, in this subclause descriptive statistics are presented on types of error that were made during different types of scenario (different services involved). Data of sixteen younger subjects - averaged over conditions - were used. From the error taxonomy that was defined in subclause 7.4 a number of types of error were constructed (table 22).

Table 22: The categories that were used to analyse types of error

Error type	Description
A	No observable effort was made to activate a service
B	Wrong service number
C	Wrong service number, but right category
D	Wrong moment of (de-)activating the service
E	Totally wrong syntax
F	Deactivated instead of activated, or vice versa
G	First button OK (i.e. * or # correct)

Figures 17 to 27 show the different types of error for each service that was used in the experiment.

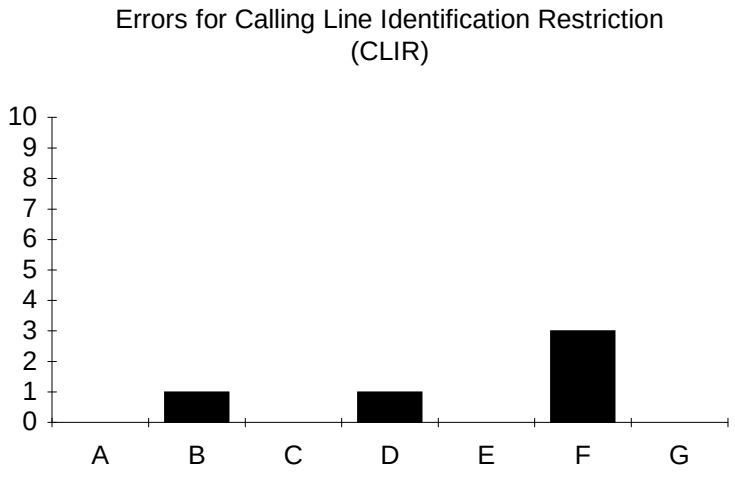

Figure 17: Number and types of error that were made when subjects were supposed to activate the CLIR service

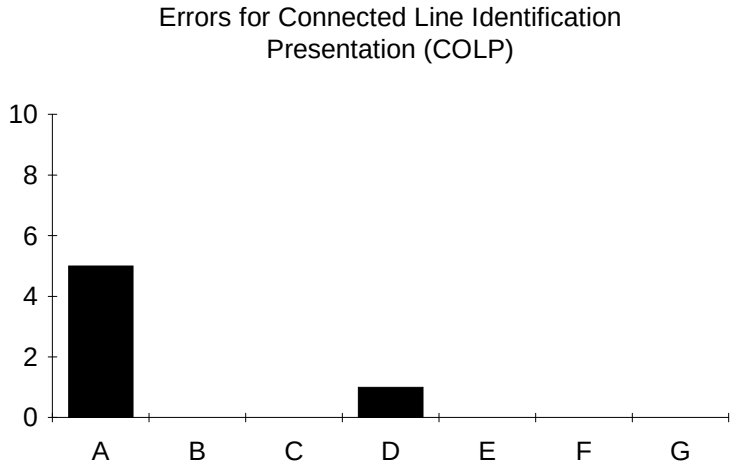

Figure 18: Number and types of error that were made when subjects were supposed to activate the COLP service

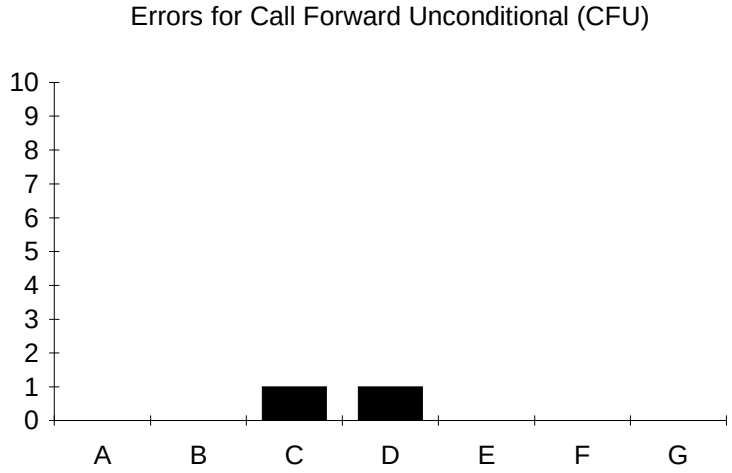

Figure 19: Number and types of error that were made when subjects were supposed to activate the CFU service

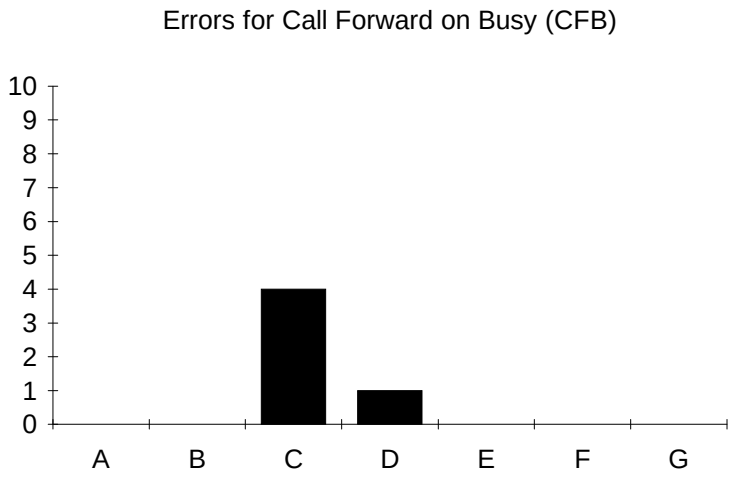

Figure 20: Number and types of error that were made when subjects were supposed to activate the CFB service

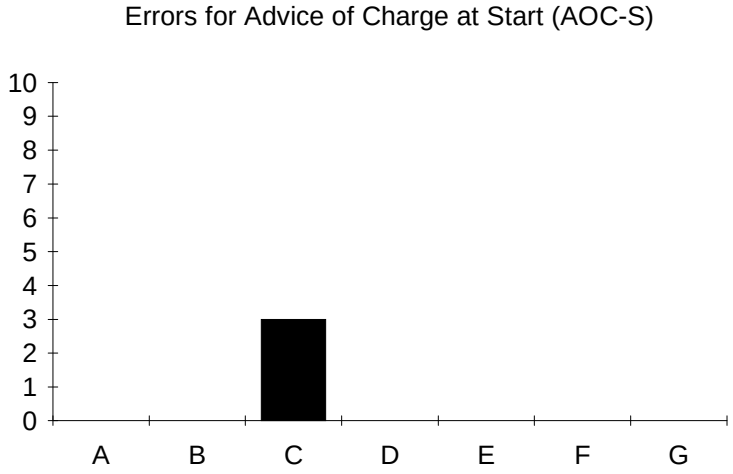

Figure 21: Number and types of error that were made when subjects were supposed to activate the AOC-S service

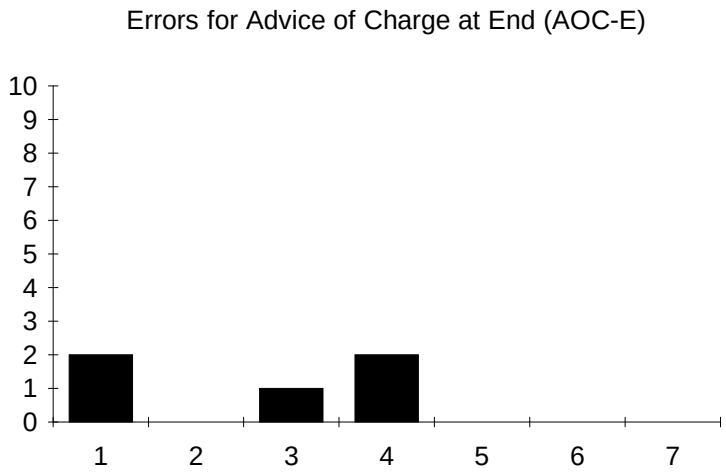

Figure 22: Number and types of error that were made when subjects were supposed to activate the AOC-E service

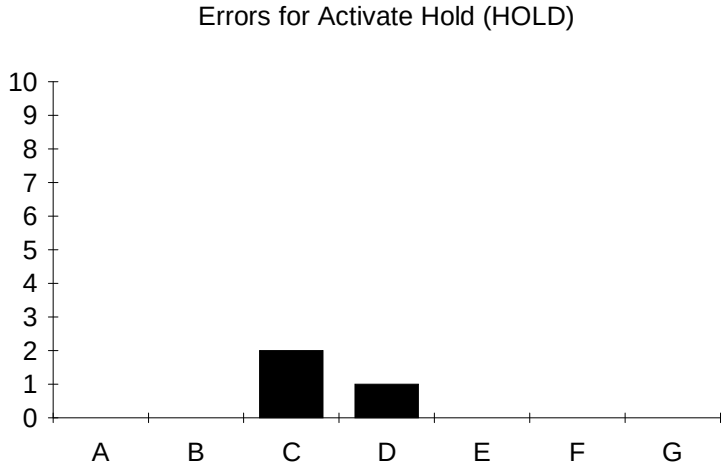

Figure 23: Number and types of error that were made when subjects were supposed to activate HOLD in the activate/deactivate HOLD scenario

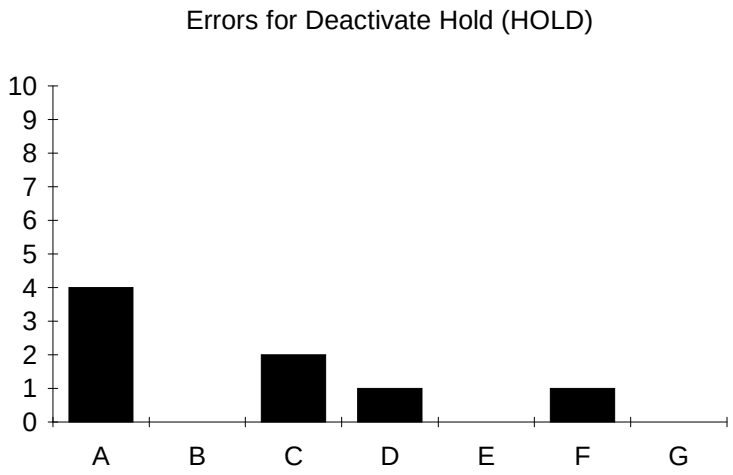

Figure 24: Number and types of error that were made when subjects were supposed to deactivate HOLD in the activate/deactivate HOLD scenario

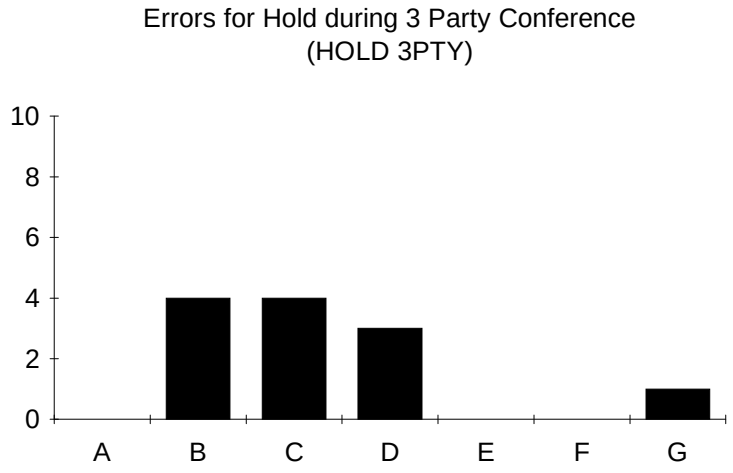

Figure 25: Number and types of error that were made when subjects were supposed to activate HOLD in the 3PTY scenario

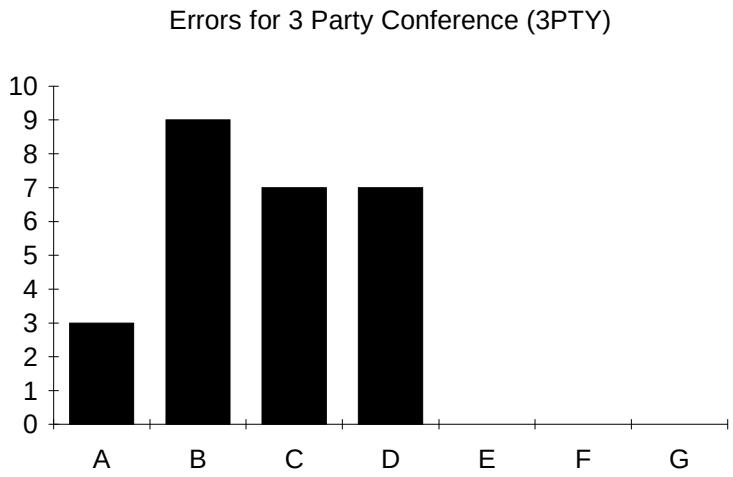

Figure 26: Number and types of error that were made when subjects were supposed to activate the 3PTY service

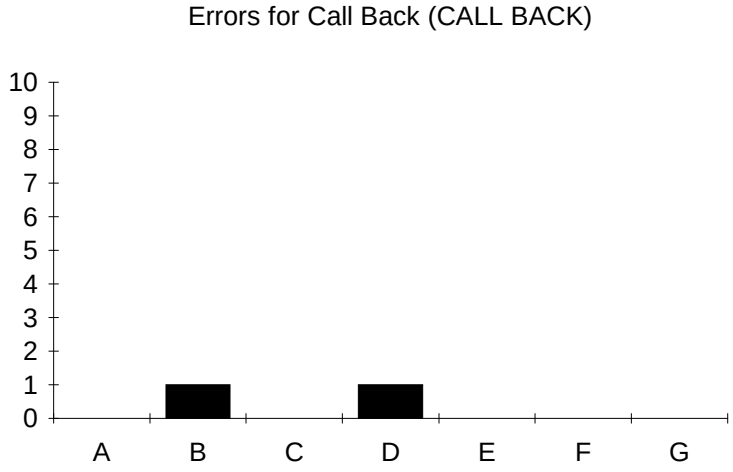

Figure 27: Number and types of error that were made when subjects were supposed to activate the "automatic call-back" service

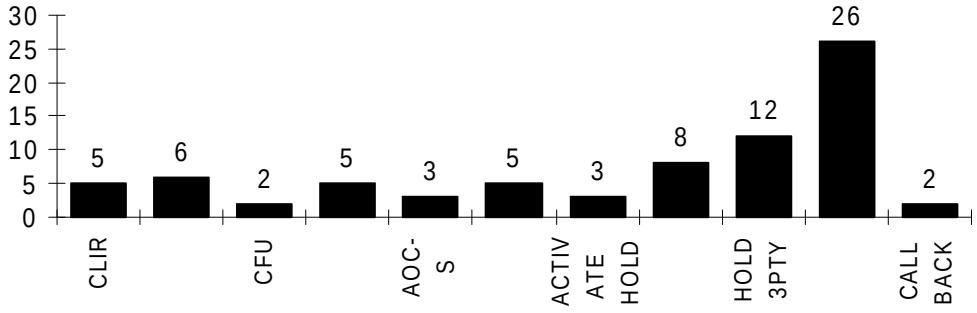

Figure 28: Total number of errors for each type of service that was used in the experiment

Figure 28 shows the total number of errors for each service type that was used in the experiment. The services are (from left to right): CLIR, COLP, CFU, CFB, AOC-S, AOC-E, ACTIVATE HOLD, DEACTIVATE HOLD, HOLD 3PTY, 3PTY and "automatic" CALLBACK.

The next clause deals with a number of issues concerning the method of experimentation. Furthermore, the results that were presented in the present clause are discussed.

9 Discussion

As stated in the preface, explorative research generates more questions than it answers. The same holds for parts of the current study. The first part of this clause deals with some issues concerning the method of experimentation. In the second part the results that were presented in the previous clause are discussed.

9.1 Discussion of the Experimental Method

9.1.1 Use of scenarios

In this study subjects solved "real life" problems. They were guided by no means, except for the User Manual, Table of Services and the instructions and feedback in the user interface. Instead of some other studies where subjects are told what service to activate, in the current experiment they had to find out themselves what the appropriate functionality was to establish the goal state.

As found many times before, the use of scenarios may be a very valid way to study the usability of services. However, it is very difficult to collect enough data for statistical analyses, because subjects take relatively long to complete one scenario. For instance, in the current experiment there were only 25 scenarios. This is quite little compared to other experimental studies where subjects sometimes complete a few hundred of trials. A solution to this problem might be to use more subjects. However, this was not possible within the scope of this study and probably will not be for a lot of other usability studies as well. As a consequence of the small amount of data statistical power was quite low.

9.1.2 "Framing"

A second important aspect of scenarios is their "framing" or wording. Changing one word in a scenario may change the subject's perception of the task goal that is set entirely. Moreover, in the pilot study it proved to be very difficult to construct scenarios that were clear and unambiguous to everyone. The same goes for User Manuals, Tables of Services and texts in the user interface. For example, in the CAT condition CLIP/CLIR had one service code (10). This means that both services were combined into one single service, called *Show number to destination*. Typing *10* would activate *Show number to destination* (CLIP), typing #10# would deactivate it (CLIR). If a subject had typed #10, the interface text had to be either "Show number to destination", the auditory instruction being "If you want to deactivate this service ..." or "Do not show number to destination" the auditory instruction being "If you want to deactivate this service ... " also. In the pilot study subjects were confused by the text "Show number to destination", because they did not want to show their number (they did not hear the auditory instruction until after a 2,5 seconds and first saw only the text on the display). In the latter case they were confused by the combination of visual and auditory texts resulting in something like "deactivate a service which prevents your number being shown to the destination". The CLIP/CLIR combination is inherently difficult anyway, because according to ETSI descriptions either CLIP is active or CLIR is active. It is not clear whether one should activate CLIR or deactivate CLIP and vice versa. In this study the services were presented in such a way that CLIP is the active state of the "Show number to destination" service, and CLIR is the inactive. Another problem turned out to be explaining the concept "default". In ISDN the user can subscribe to either default CLIP or default CLIR. It was difficult to explain to computer naive people in a few words what default means. This finding is in line with others (Waern, 1989; p.251). The problem was solved for the moment by leaving the information from the service description and saying for each CLIR scenario "Normally the person you are calling up can see who you are."

9.1.3 Scenario button

Another issue in the use of scenarios is the indication of the start of the next scenario. This indication is necessary to be able to calculate the duration of the scenario, but also to be able to produce the right sounds. In the implementation of the user interface, scenario specific sounds were coupled to the scenario number. Therefore, the scenario number had to be raised at the start of each new scenario in order for the subject to hear the right sounds. However, the choice of the kind of implementation described in clause 6 made it impossible to let the scenario number be raised by the experimenter or entirely automatically. As a solution a scenario button was constructed, which had to be pressed every time the subject started with a new scenario. As was to be expected, subjects incidentally forgot to press the scenario button. Sometimes, upon discovering, they just pressed it until the scenario number corresponded to the scenario they were involved in. Sometimes they started doing the one or two scenarios they had already completed once more. For that reason the end of the scenario was defined as being the moment the subject completed the right commands for the first time.

9.1.4 Written information

Subjects had to read a lot in this experiment. First they had to read the User Manual, next the Table of Services. Some subjects may have been hampered by the fact they had to process a lot of written information, contributing to the large individual differences that were found. Future research employing scenarios might include a test for the estimation of subject's reading speed.

9.1.5 External validity of the learning environment

It is not likely that all of the supplementary services that were used will be introduced at once. There will be a migration path from the current situation to the one where actually a large number of services will be available. Users will be confronted with new service commands one by one which makes learning much easier. In other words, the subjects in the current study had a disadvantage. On the other hand, having an overview of all services, which illustrates the intended conceptual model in a most clear way, may have been an additional aid in constructing a mental representation of the command language, resulting in a stronger effect of learning and better retention of services commands. It is acknowledged that the experiment is not entirely representative in this way, although it is very hard to do representative research on learning within a context like the current one and a time period as short as the one that was available.

9.1.6 3PTY descriptions in the Table of Services

The description of the Three Party calling service contained a reference to the description of call HOLD. In the CAT condition this reference was direct (it referred to the service code), as the NoCAT manual contained an indirect reference (referring to the service name only). As a consequence there was a disadvantage for the NoCAT subjects, especially the elderly, because this way the task placed quite a burden on short term memory.

9.1.7 Implementation

As each scenario consisted of a number of stages, the HyperCard® stack was programmed in such a way that the user would proceed to a next stage once he had (de-)activated a service. A disadvantage of this way of implementation was that (de-)activation of services that were equivalent to the intended service resulted in a successful progress to the next stage as well. For instance, in a certain stage the subject could activate 3PTY as well as ECT, resulting in the same sounds to be presented by the system. The use of this fixed order of stages (sounds) in each scenario made it impossible for a subject to return to a previous stage and try again. Once the subject had made an error the entire scenario would go wrong. This means that subjects received a severe penalty even for minor mistakes. The second consequence was that subjects did get feedback on whether they had activated a service and whether *that* service had been activated in the right way, but they did not get conclusive feedback on whether the right service had been activated, unless they did something that caused the application to produce the wrong sounds. Some subjects in fact complained about the fact that they did not know whether they had acted in the right way and were ready to proceed to the next scenario. Unfortunately there was no other solution possible, except for perhaps a "Wizard of Oz" technique (whereby a human operator interprets user actions and translates them into system states. This way a user interface can be simulated (e.g. speech recognizing input device)). For practical purposes it was necessary to automate the experiment as much as possible.

Besides, it was considered appropriate to let every subject try once, because this way lengthy retrying of subjects and large differences in speed-accuracy trade-off were avoided. However, this combination of lack of feedback on and punishment for certain errors may explain the fact that the effect of practice is less for number of errors than for duration of scenarios and duration of manual consultations.

9.1.8 The user interface

There are a few aspects of the user interface itself which are worth mentioning. Firstly, it turned out to be very hard to construct texts that were both short enough (2 x 20 characters) and understandable. As said before, some texts were confusing to some subjects and not to others. None of the texts was confusing to all subjects. A final example of a text which may not have been immediately clear was the auditory instruction "If you want to proceed type ..; if you want to choose another ...". By choosing any of the menu options the subject would in fact proceed. So, in other words it was not possible *not* to proceed. Consequently some subjects were confused, because they saw choosing another service or group as a way of proceeding too. A better text would have been "Press the number of the preferred group, or type .. to choose another service/group.". Despite these (minor) flaws the interface as a whole seemed to be intelligible to most of the younger subjects.

9.1.9 Use of * and

With regard to the use of * and #, the following trade-off was an issue. On the one hand the consistent use of the same symbol for prefix, separator and suffix in commands would facilitate the user in remembering it. On the other hand, * and/or # had to be used as some kind of "escape" button as well (in case of choosing the wrong category or service the user can go back and choose another). It is commonly known in the field of human factors that for this kind of function there should always be the same key. However, it was not possible to satisfy both criteria. Either one of the two options shown in table 23 was available. Option A was chosen, partly for reasons of implementation, partly because this syntax structure was expected to be better remembered which was an important issue in this study. Either the syntax was easy to use for those who mostly typed a correct command at once (option A) or the syntax was easy to use for those who made a lot of use of the "go back" function (B).

Table 23: Two options for a command syntax

option A		**option B**	
.......*	go back with #	*..*.....*	go back with #
#..#.....#	go back with *	#..*.....*	go back with #

9.2 Discussion of Results

9.2.1 Subjects who did not finish the experiment

One of the salient findings of this study is the fact that so many (elderly) subjects did not finish the experiment and that the others had a mean error percentage of 17,2 %. It seemed that the former did manage to activate the right service every now and then. However, they took such a long time doing so that they exceeded the time limit. This is illustrated by the fact that a significant age difference in mean duration of scenarios was found. The fact that three elderly males did perform equally well or even better than the younger subjects and that three of the younger subjects did not complete the experiment indicates that there may be another factor involved. It seems that it is not necessarily age which determines a subject's ability to use the services. The elderly subjects who did finish the experiment had moderate to high education and SPM scores. The younger subjects who did not solve all scenarios had relatively low education and SPM scores. Therefore, it must be concluded that education, occupation, SPM score and perhaps some construct as "lifestyle" may be important determinants of performance. One could argue that the stereotype of the poorly educated elderly housewife not knowing how to operate devices is confirmed once again, but this time it was found that there are also younger (and male!) people who encounter the same problems.

9.2.2 Effect of practice

From the data presented in the previous clause it must be concluded that there is an effect of practice with the services on all performance measures. There was a significant gain in the speed of completing a scenario with time. Subjects spent increasingly less time consulting the user manual. They made also progressively fewer errors. This effect of practice is experienced by the subjects themselves as well, because they indicated that they had invested less effort in the second part of the experiment than in the first.

Care must taken in interpreting these effects as being solely the result of learning how to operate the supplementary services. In the beginning of the experiment subjects have to get acquainted with the task environment. They have to learn "how to do a scenario". After a while this procedural knowledge may have become automated, resulting in a higher speed of completing the scenarios. These two effects of practice can not be separated. With respect to the duration of manual consultations there are two possible explanations of the change in performance. The first is that subjects learnt how to match task goals and service functionality. After practising they may have used the manual for verifying solutions instead of searching them, actively remembering the locations of relevant services. A second possibility is that they have learnt how to search the manual. Although subjects were instructed to use the manual as little as possible, they hardly managed to solve scenarios without referring to it. This means that their representations of the services and the command language were not so detailed that they allowed inference of operating procedures on the basis of the representation alone. In this experiment the information retrieval model was a copy of the conceptual model of the command language. Instead of using the conceptual model of the command language to remember service codes syntax, subjects may have learnt how to use it to search the manual. The first explanation suggests that subjects simply remembered the services they used and consequently needed the manual less. The other explanation supports the hypothesis that the subjects incorporated the conceptual model during the task at least to some extent and that they used it to search the manual. The fact that subjects made fewer errors with practice shows that they may have understood the functionality and operating commands of services progressively better.

9.2.3 Effects of transmitted conceptual model

There were no effects of type of transmitted conceptual model on total duration of task completion and total number of errors. There was an effect on total duration of manual consultations. The same explanations serve as for the learning effect described above. On the one hand subjects in the CAT condition may indeed have remembered the operation procedures better than subjects in the NoCAT condition. They may also have had a "better guess" of where to find the description of the service on the basis of the conceptual model they incorporated. An alternative explanation is that the structure of the CAT manual facilitated far more efficient search strategies than that of the NoCAT manual. In the latter case, it is not the conceptual model that explains the difference, but the organization of the manual (which reflects the model). In order to find out which of the two explanations is most probable (whether the conceptual model did facilitate performance) it is necessary to determine performance of subjects when they have no user manual available (see discussion of model tests).

CAT subjects used less time consulting the user manual, but the total amount of time they took to complete all scenarios was equal to that of NoCAT subjects. This is rather peculiar unless they used the structure of the user interface to search the preferred service, which was in fact observed on a number of occasions. For instance, it was possible to employ the following search procedure (it was in fact one of the design principles):

- type a * and then a digit (group number);
- check whether the group name corresponds to the group you are looking for or to the task goal;
- if it does, proceed by typing another digit (service number);
- if it does not, type # to remove the group number and type another digit; then repeat the procedure described above.

CAT subjects may have used this procedure more often than NoCAT subjects resulting in a reduction of time needed to consult the manual, but in an equal total duration of task completion. Furthermore, a secondary data analysis might reveal whether CAT and NoCAT made the same *type* of errors.

It must be concluded that there were no differences in learning curves between CAT and NoCAT on any of the performance measures. Therefore, the hypothesis that states that CAT subjects learn faster and longer (they acquire a higher level of skill) is rejected.

9.2.4 Relation between SPM score and performance

Although a relation between SPM score and performance was to be expected (smarter people perform better), the correlation that emerged for total duration and total number of errors was surprisingly strong (Pearson $r \pm -.5$). In addition the variation in SPM scores within the sample was large (within the same condition one subject's score was in the 0,99 percentile, another one in the 0,10 percentile). As a result, within group variation in performance was also quite large, which may have overruled some of the probably very subtle effects of different conceptual models.

It is remarkable that there is no correlation between SPM score and duration of manual consultations. The SPM measures the ability to find a "solution" to an abstract situation, which is comparable to finding the right service that matches the (abstract) task goal. Apparently subjects with a lower SPM score did not feel the need to search longer in the manual, although they performed worse than subjects with a high SPM score, with regard to total duration and total number of errors. They may have relied on trial and error behaviour. This does agree with the strong correlation that was found between SPM score and total number of errors.

9.2.5 Remembering the service commands

A significant difference was found between CAT subjects' ability to remember service codes and operating procedures and that of NoCAT subjects. As the only difference between both conditions was the explicitly offered model in CAT condition, it is possible to conclude that the subjects managed to incorporated the conceptual model and used it to answer the questions. The CAT model test contained additional questions on group numbers and names. An alternative explanation is therefore, that remembering the questions on grouping allowed them to reconstruct the model which lead to a better recall of codes and procedures.

The subjects who returned the second model test remembered the services commands quite well. There was a 21 % decline in recall of services over two weeks, averaged over type of transmitted conceptual model, which means that the consistent structure of the language may in fact have facilitated the subjects in remembering the commands.

9.2.6 Understanding the functionality of a service

The difference between Call Forwarding Unconditional and Call Forwarding when Busy is fairly small. Several subjects used CFB instead of CFU, because they reasoned "As a long as I am still at home I want to be available for incoming calls". This indicates that they are able to understand and use the differences in functionality. Another illustration of their understanding of the concepts emerged from the behaviour of more than half of the subjects in the "truant" scenario. The goal was to make sure that anyone who called the office phone number would reach the subject at home. This means simply activating CFU. However, many subjects started by activating CLIR to make sure that no one could see to what phone number their call had been forwarded. This illustrates the fact that they did manage to develop and use some rudimentary network concepts.

As a counter example of this observation several subjects solved the three party scenarios by putting the first call on hold, calling the third party and then getting the first call from hold. According to the description of HOLD in the Tables of Services this would result in interchanging the two phone calls, i.e. switching from hold to active respectively active to hold. However, it had not been possible to implement this. The effect of subjects' actions was that they indeed did establish a three party conversation. As their solution of the scenario was rewarded by the fact that the scenario proceeded the way it should, these subjects did not learn, but were encouraged to maintain their "waiting room" metaphor, as one of them expressed it.

9.2.7 Types of error

Due to a time limit it was not possible to analyse the types of error extensively. The most important observation is that subjects made more errors on trying to establish a three party call or putting a call on hold. This agrees with the mental model explanation, because operating these services requires more complex mental representations with regard to structure as well as dynamics. A secondary data analysis might reveal more subtle differences.

The next clause contains a summary of the main conclusions from the discussion above. Furthermore, a number of recommendations and possible directions for further study are discussed.

10 Conclusions and Recommendations

10.1 Conclusions

The discussion in the previous clause can be summarized by a number of conclusions based on statistical effects that were found as well as a number of observations that were made.

With regard to age-related differences in performance conclusions are the following:

- For certain people the supplementary services are too difficult;

- It is not ageing per se which determines whether someone will understand and learn how to use the services. There are also younger people for whom the services are too difficult. These people share a number of characteristics which have to do with level of education and the ability to use abstract concepts in their thinking as well as experience with information technology;

- Most younger people can learn how to use the services. In most cases they manage to understand the functionality of the services involved, but not always (e.g. 3PTY).

With regard to the presentation of an explicit conceptual model and effects of practice that were found in the younger subgroup the conclusions can be summarized as follows:

- Performance both in terms of speed and errors will improve with practice;

- Offering an explicitly structured conceptual model of the command language (code scheme) facilitates users in remembering the operation of supplementary services. It also reduces the time needed to consult the manual;

- Having a mental representation that reflects this model reduces neither time needed for operating the services, nor number of errors It does also not lead to faster learning of the command language. Finally, subjective mental effort is not reduced.

The next subclause deals with a number of recommendations based on the conclusions above. Furthermore, a number of possible future research directions are stated with regard to the usability of supplementary services as well as cognitive psychological research on ageing and mental models.

10.2 Recommendations and Further Study

10.2.1 Account for subgroups

The most important recommendation regards inter individual or rather inter subgroup differences. On the one hand service providers should be very careful in determining the population of interest for a certain service. If a service is offered on the public telephony network and all telephone users are seriously considered to be the target group, the service provider should ensure that the service is intelligible for everyone by employing "good" human factors and not just keep fingers crossed and hope for a lot of customers. It is in fact possible to offer understandable services, but this requires a lot of usability research, which is usually considered too time consuming. On the other hand, the group of users that would be reached in case of more understandable services is large enough to make it worth while to do fundamental human factors research on different service types, instead of investigating the usability of single services. Usability researchers of single services should at least be aware of the large individual differences and make sure that all subgroups of the population of interest are represented in the sample. Finally, it is important to note that the problem of the current elderly population is not a temporary one. There will always be computer naive people who have trouble using information technology.

10.2.2 CLIP/CLIR

With regard to CLIP/CLIR it would be useful to do another experiment to compare different ways of explaining the default concept and see whether the services should be present as two forms of one new service, rather than two separate services. It is stressed here that the service might seem to be very simple at first sight, but that it becomes in fact quite difficult as soon as the "default" concept is introduced.

10.2.3 Three party calling

The supplementary service combination HOLD-3PTY is difficult to use. A number of subjects in this study thought they had to get the held call from the waiting room in order to be able to talk to them. The thought they could build their own three party conversation instead of using the special three party supplementary service. Therefore it is necessary to determine very carefully how this service should be presented to the public. Using the aforementioned metaphor is an option. Further study on this service is indicated in particular.

10.2.4 Auditory feedback

A future experimental study might investigate the possibility of employing auditory feedback instead of visual and omit the display. If people are able to operate such kind of phone based interface in the same way as in the present study, the minimal interface would be a PBI+ rather than PBI++. Instructions and feedback should be given by two clearly discernible voices (e.g. high/low or male/female voices). In case of such an auditory interface a solution would have to be found for the fact that visual feedback is displayed semi-permanently on the screen as opposed to volatile auditory information.

10.2.5 Conceptual models of the functionality of services

As argued in clause 4, users' mental representations of the functionality of services are very important. Therefore, future research might aim at identifying candidate conceptual models by examining and interviewing intermittent knowledgeable users, for instance by means of think-aloud protocols. The conceptual models that are found might be used to construct comprehensible user manuals.

10.2.6 Avoid written scenario descriptions

In future experiments employing scenarios researchers should take care that they use as little written information as possible. Rather, it would be wise to describe the initial situation by means of short video story or a "comic strip". This way not only an initial disadvantage for the lower educated is avoided. The other subjects are relieved from processing a host of written information as well.

10.2.7 Future research on age-related differences

Now that it has been found that age per se is probably not the only determining factor, future research should aim at identifying what the precise factors are and how elderly can be aided in understanding information technology. However, the focus should not be solely on the elderly, but also on younger people who have problems using information technology. Instead of directing attention to the elderly as a separate group and neglecting the lower educated and computer naive younger people - who seem to have comparable problems - both could be grouped together and should be of joined interest.

10.2.8 Research on conceptual models

Future usability studies on how to communicate a conceptual model to "the" user might employ a very homogeneous sample, include SPM score as a between subjects factor or match subjects explicitly on their SPM scores. In a follow up study factorial analyses of educational and occupational determinants of performance might also be performed.

All in all it must be concluded that (ISDN) supplementary service-like features can be made usable at least for people with some level of affiliation to the use of information technology or other activities that employ abstract concepts. Service providers should be aware of the widely varying characteristics and abilities of the various subgroups in their target populations. On the one hand the services might simply be introduced, regardless of their usability. On the other hand, service providers could choose to accept the challenge to try to make the services usable to a heterogeneous population. By doing so they might open up a large consumer market.

History

Document history	
October 1996	First Edition

ISBN 2-7437-0411-X - Edition complète
ISBN 2-7437-0864-6 - Partie 4
Dépôt légal : Octobre 1996