The Role of Individual Differences in Dialogue Engineering for Automated Telephone Services

Stephen Love

Doctor of Philosophy

University of Edinburgh

1997

Declaration

I hereby declare that the work presented in this thesis is my own, unless explicitly indicated otherwise. I further declare that this thesis has not been submitted in this or in any other form to this or any other university.

Acknowledgements

Many people have helped in a variety of ways in the generation of this text. I am very grateful to all of them.

I would like to thank my supervisors, Mervyn Jack and John Foster, who have been patient and helpful. I would also like to thank my colleagues on the Intelligent Dialogues Project for the help they gave me with different aspects of the thesis. My thanks also go to Fred Stentiford from BT Laboratories for his support. I would also like to thank Safaa Fathy for the advice she gave me on referencing.

Abstract

The research work reported in this thesis examines the hypothesis that users' individual differences affect performance on computer based tasks, specifically in the context of automated telephone services. The experimental framework was based on a three stage approach for identifying, isolating and accommodating individual characteristics which have a significant effect on user performance when interacting with a computer system.

The first experiment reported here addressed individual characteristics which have a significant effect on users' performance with an automated catalogue service using three different input modes: isolated word, connected word, keypad entry. The results obtained from this experiment indicate that in terms of identifying individual differences, the salient individual differences for keypad and connected word performance (measured by call duration) on the automated catalogue service are age, verbal ability and spatial ability. No significant results were obtained for subjects' performance on the isolated word version of the service.

The second experiment reported here extends the work of the first experiment to a more complex task (a hierarchically structured automated music catalogue) and isolates where these salient individual characteristics have their most significant effects. The results obtained in this experiment confirm the identify stage of the first experiment by showing significant differences in performance between age groups and between low and high verbal ability subjects. The results also showed significant differences in performance between these groups of subjects when the task was broken down into three component processes.

Having identified and isolated these salient individual characteristics the final stage in the work was to accommodate for the needs of this group of users. This was achieved by comparing the performance of subjects using speech input and keypad entry versions of the automated music catalogue. The results obtained from this experiment, indicate that speech was partially successful in accommodating older and low verbal ability subjects.

Chapter 1: Introduction

Chapter 2: Individual Differences in Human-Computer Interaction

Chapter 3: Methodological Background

Chapter 4: Assaying Individual Differences

Chapter 5: Isolating Individual Differences

Chapter 6: Accommodating Individual Differences

Chapter 1: Introduction

The telephone, the telephone network and automated telephone services provide the first point of contact for human-machine communication for a growing number of people. Public access to advanced telephony services is currently significantly higher than access to other forms of information technology products such as personal computers and so the telephone is, and will increasingly become, a vital tool in accessing information and services such as home banking or home shopping from the comfort of the home or from the office. It is important therefore that the engineering design of automated telephone services, based on spoken language user interfaces or phone-based interfaces (PBI's) as they are also known (Halstead-Nusslock, 1989), should focus on the creation of effective, efficient and satisfying interfaces which take into consideration the needs of various types of users. The aim of this thesis is to identify an inventory of user characteristics, here termed *individual differences* which have a significant effect on users' performance and attitude when using phone-based interfaces with spoken language dialogue systems and to investigate methods to accommodate important individual differences in the system engineering design of automated telephone services in order to improve the usability of these types of system.

A telephone-based user interface has a fundamentally different structure to a screen-based user interface. With a screen-based interface, information can be displayed by means of a variety of different media where the user can look at screen graphics, icons, animation and text. Importantly, these various media can be deployed in the user interface concurrently. In contrast, the telephone-based interface can offer information only in serial fashion with the user responding to sequences of auditory messages from the system. The inherently linear nature of a telephone-based spoken dialogue results in specific limitations. For example, the user does not have random access capability for menu selection which is available to the user of a graphical user interface. In the case of a screen-based interface the user can exploit short-term memory aids displayed on screen and these can be processed repeatedly. A telephone-based user interface on the other hand does not benefit from any such short-term memory aids. Other memory aids available to the screen-based user include the flashing of the cursor, together with menus and log files. The telephone-based user has to rely on the content of the current system message and, perhaps, on overall familiarity with the system structure. Therefore although telephones are readily available to a large number of people, there are limitations which have to be taken into account when it comes to designing telephone-based systems. Designers have tackled these usability problems in a variety of ways.

One aspect of the usability of automated telephone services which has received considerable attention is the mode of communication between the user and the system. There are a growing number of automated telephone information services which use speech input and/or output for applications such as railway timetable information and home banking services. However, there are automated telephone service design issues which remain to be addressed by designers in this field. When it was first introduced, user acceptance of systems using synthetic speech output for automated information systems, was high (Gould and Boies, 1984). However, as telephone-based information systems became more prevalent it became apparent that, in many cases, that these systems had been implemented without an adequate assessment of their usability. For example, Martin, Williges and Williges (1992) in their overview paper on design guidelines for telephone-based information systems, state that users often complain of poorly organised or unnecessary system messages. In the move towards developing more user-friendly systems, increased emphasis was placed on behavioural data which focused on performance measures such as task completion times and error rates in order to provide information on the appropriate style, structure and content of human-computer dialogues for different types of applications.

1.1. Automated Telephone Service Design Issues

Halstead-Nusslock (1989) suggests that automated telephone service designers should focus on two aspects of the user interface. These are firstly the mode of communication and secondly the dialogue structure. With respect to mode of communication, the first task facing the designer is whether to allow the user to respond by speech or push button keypad input or both. Halstead-Nusslock suggests that designers should consider speech input when the users' hands or eyes are involved in a simultaneous task and the active vocabulary that the system will be required to recognise is relatively small. This is necessary to improve the speech recognition accuracy of the system. Halstead-Nusslock recommends the use of keypad entry in dialogues because error rates when using keypad input are orders of magnitude lower than the best speech recognition systems.

Fay (1993) also analysed whether users prefer keypad or automatic speech recognition input, to determine if the limitations of automatic speech recognition (such as recognition errors and limited vocabulary input from the user) make it so difficult to use that people would preferentially use the keypad mode. In addition, that investigation attempted to determine if users have similar attitudes to automatic speech recognition and keypad systems. In Fay's experiment the subjects were given two scenarios, each of which had to be completed using both speech and keypad input. The first scenario involved the subjects registering for college courses (the subjects were all American college students) and the second scenario involved the subjects contacting a telephone company in order to arrange to have their telephone disconnected.

The results showed that there was no simple answer as to whether or not people prefer automatic speech recognition or keypad input, concluding that user preference is dependent on how the technology is employed. In experiments involving recognition of college course details, there was a strong preference for keypad input. However, in experiments where the task was to make arrangements to have a telephone disconnected, the preference was for automatic speech recognition. Fay did not offer reasons for these findings and questions arising from the findings form the point of departure for this thesis where Chapter Four and Chapter Six explore the possibility that preference for speech interface or keypad interface is related to user individual differences such as personality and spatial or verbal ability, age and gender.

Dialogue structure has an important role to play in system design. In a person-to-person telephone conversation both parties have roles and responsibilities in maintaining the dialogue flow. These are mainly implicit and are learned via a process of social experience. In contrast, for automated dialogues, Halstead-Nusslock (1989) suggests that there are two main features to be considered in designing the role to be played in a dialogue by the system. The first feature concerns **system prompts**, which can be defined as utterances made by the system which present the user with a list of legal choices or actions which can be taken.

The second feature concerns the **system messages** which are utterances made by the system to inform the user of the current status of their on-going interaction.

Halstead-Nusslock (1989) found that auditory menus make effective system prompts since all of the information that users require in order to progress in their interaction is made available. These results suggest that system prompts as auditory menus need to exhibit three components in order to be maximally effective. First, system prompts should have a *title* or *heading* to provide immediate feedback to users about their location in the system and to act as a form of localised priming. Secondly, system prompts should provide users with a *list of options,* which should contain selections like "in order to do X, enter Y". This 'goal-action' format is recommended in the belief that it follows the same logical sequence used when individuals process information. This format also minimises the cognitive load placed on short-term memory (which has a limited storage capacity). Thirdly, all system prompts should have an *ending* which indicates the termination of the menu and could contain instructions on what to do next. Halstead-Nusslock (1989) claims that auditory menus are the best initial choice for system prompts, and recommends that system prompts should be interruptable. In addition, for expert users, the interface designer should provide another, optional dialogue, without menus or system prompts to allow rapid progress through the service (this is also known as a **fast track dialogue**).

Linked to this is the design of **system messages**. Halstead-Nusslock (1989) suggests that there should be three types used in telephone-based interfaces:

- *Error messages* should identify a system or user error and how it might be corrected.

- *Completion messages* should give users feedback on a successfully completed step in their on-going interaction.

- *Working messages* which provide information to users about work in progress. For example, a work in progress message could tell the user how long it will take the system to complete the particular action they have just initiated.

The user's role in the dialogue involves two types of utterance. The first involves *selection* with the user indicating the item they wish to select from a list of alternatives. The choices offered to the user at any one time need to be limited and each option needs to be easily distinguished in auditory and semantic terms from all the others that are available at the same time. The second type of utterance the user can make concerns *controls* which allow the user to navigate through a dialogue. In telephone-based interactions users need to be able to direct the dialogue flow in an efficient manner by means of the controls that are available to them. For example, where controls are assigned to specific keys on the telephone keypad it is crucial that these key assignments should be kept consistent throughout the dialogue. For example, if the star "*" key is used at the first level of a hierarchically structured dialogue to allow the user to leave the service, this key/function should still be the case at all levels of the dialogue structure. In general, keypad telephone interfaces need to implement control functions on keys which allow:

- interruption - where the user can interrupt a prompt or message from the system.

- repetition - when the user can request the system to repeat the most recent prompt or message.

- access to help - where the user is in some difficulty and needs assistance from the system.

Overall, in a dialogue design each set of options and controls should be complete, robust and oriented towards the task the user has to complete.

Choinere, Robert and Descout (1991) also propose design guidelines for speech recognition-based automated telephone services devised from user responses to an automated telephone-receptionist application. They stated that in order to have satisfactory speech recognition performance over the telephone line the user interface must ensure that the user's needs and expectations of the system are matched with its actual performance and functional capabilities. In addition they pointed out that most work has focused on evaluating only the speech recognition performance, e.g. Wilpon (1989) and Dutiot (1987), and as a result, system designers have been left without proper user interface design recommendations. A maxim often quoted in interface design is "know the user" (Hansen, 1971). Choinere, Robert and Descout (1991) note how this is often ignored by designers of applications that are designed to be used by the general public. They argue that with a good understanding of users, the designer can engineer the system functionality and options to address the needs of the users. As for their own recommendations for user interface design they emphasise the need for *interruptability*, an aspect of system design which allows the user to interrupt a system prompt at any time. This allows experienced users to move quickly through the system while allowing novice users to listen to all the information given in a dialogue prompt from the system before making a choice about what to do next. They also argue that on-line help should be adaptive. That is, it should "evolve" with the dialogue. A system which provides good on-line help should provide users with information about where they are in a dialogue and what options are currently available to them.

Dialogue structure issues are implicitly linked to mode of communication issues in telephone-based interface design. Automatic speech recognition systems potentially offer a more human interface than keypad systems because speaking into a telephone is considered more 'natural' than pressing the buttons on the keypad. However, the style of conversation that a user can have with an automated speech recognition system is not currently as linguistically rich as in a human-human conversation. In an automated telephone service users have only a limited vocabulary at their disposal, in order to keep speech recognition errors made by the system to a minimum. Jones, Hapeshi and Frankish (1989) included a reference to the need for limited vocabulary in their list of design guidelines for interfaces which use speech recognition, with the idea that user expectations are better matched to the capabilities of the system, if they are advised of the specific vocabulary to use when interacting with a system.

Karis and Dobroth (1991) argued that the normal rhythm of human conversation is never really obtained in human-machine dialogues. Nickerson (1976) suggested that a good dialogue between a human and a computer should allow for 'mixed initiative' where either the user or the automated service is allowed to direct the interaction making it clear at all times which interlocutor is in control and retaining a sense of presence of the non-speaking side in the same way as the hour glass in a graphical user interface indicates to the user that the system is active. In human-human conversation, many of the cues necessary for achieving these conditions of continuity are in fact non-verbal. Furthermore, Murray and Bevan (1984) argue that in general, human behaviour is not just goal-oriented, it also has a social content and can be used to establish and communicate the relative power and status of the communicator or can be used to gain social acceptance. Social content strongly influences the communicator's perceptions of the communication and the interpersonal relationship involved. In contrast, human-machine systems, on the other hand, are generally goal-oriented and task specific, lacking the shared knowledge or experience of human-human communication. As a result, when people speak to automated systems (even to answering machines) they tend to adopt a task-oriented style often found between business partners (Egan, 1988). Dialogue styles like this may not suit, say, individuals who tend to be person oriented and need social context, which is missing from this kind of human-machine interaction. Therefore when some individuals attempt to use speech input with computers their habitual methods of communication are abnormally constrained.

There have been attempts to design telephone interface systems based on a more conversational dialogue approach. For example, Schmandt (1987) describes an automated telephone service called "The Phone Slave", a system which is a conversational answering machine, taking messages from users by asking them a series of questions and storing their messages digitally. When the system answers the telephone it immediately takes the initiative in the conversation because it has no real ability to answer questions and cannot in fact use speech recognition to try to *understand* what the user has said. Instead the system records messages from the user by asking them a series of questions like "Who's calling please?" and "What's this in reference to?". These (recorded) audio files can be subsequently accessed in sequence by the system's owner. The Phone Slave does not make any attempt to understand the content of the messages left by any of the callers. Instead, it makes a note of the sequence in which the individual message components were recorded and presents these in appropriate order to the owner. For example the expert owner who is familiar with the system operation and whose speech has been registered by the system's speech recognition processor could ask "who left messages?" and this would result in the system playing back all the responses to its query "who's calling please?". Schmandt found that the interface proved to be very effective in eliciting appropriate message components from callers, attributing the success of the conversational style to the apparent high quality of the spoken prompts provided by the system. To take a message requires co-operative behaviour and there is no reason to think that callers will not follow conventional rules. By asking a series of questions, as opposed to a message such as "leave your message after the beep", the system makes it easier for the user to leave a more complete message as a series of components. In doing so, the system maintains its ability to control the conversation and the limited discourse protects the system's limited "intelligence" from being exposed.

Peckham (1995) discusses conversational interaction between users and automated telephone systems. This type of interaction resembles human-human communication where initiative can shift from user to computer and vice-versa, misunderstandings are easily corrected and vocabulary is unconstrained. He believes that this style of interaction will be more efficient and effective by allowing users to explicitly state their goals rather than having to negotiate their way through a menu structure (using single words, phrases, continuous digits or the telephone keypad) in order to complete a task. He qualifies this to an extent by saying that although the conversational style will potentially create a more usable interface, there is a need to assess the technological requirements necessary to support this kind of interaction.

Based on earlier work Peckham (1993) has proposed a set of guidelines for conversational dialogues in telephone user interfaces which would allow spontaneous conversational interaction between the user and the system.

The first guideline suggested by Peckham (1993) was that there should be real time interaction between the system and the user. Human-human conversations occur in real time, therefore a system should be able to respond in the same way. Second, the system should allow for spontaneity as spoken human-human conversations contain hesitations and ill-formed expressions. Third, the ideal conversational interface should allow for extended interaction. Natural dialogues can extend beyond one turn (a turn refers to an epoch when one side is in control of the dialogue at a specific moment in the interaction) and they also display a sequential organisation. He also suggests that automated dialogues should be mixed-initiative allowing either the system or the user to take the initiative in directing the course of the interaction. Co-operative responses are also important. If the system does not find the information that the user has requested, it should provide an explanation of why it can not find it rather than generating a simple negative response. Finally, he believes that conversational dialogues should have effective error recovery (repair) strategies. For example, if the system has difficulty in interpreting a user's utterance, it should attempt to identify what the problem is and take steps to correct it in order to allow the dialogue to continue. Once systems have been developed which incorporate these capabilities in their design, Peckham believes the public will be more ready to accept automated telephone services.

This underlying assumption of Peckham's that speech input will provide the most effective interface for automated telephone services mirrors the general view held by speech technology researchers. There is a repeated claim in much of the work in speech recognition which suggests that speech input will be used whenever it is possible because it is the most 'natural' form of communication (Lea, 1980). Damper (1993) notes, however, there are differences between human-human communication and human-computer communication which imply that the advantages found in human-human communication do not necessarily transfer to human-computer interaction.

At present, telephone speech recognition systems are generally constrained in terms of user input which makes the goal of a "natural" dialogue between human and computers seem a long way off. The current technological limitations of these systems impose a burden on the user which can limit the usability of the application.

Speech recognition systems generally fall into one of two categories: speaker-dependent and speaker-independent. Speaker-dependent systems require the user to speak a sample of the words they will use when interacting with the system through a process known as **enrolment** which requires the user to speak (sometimes several times) the words in the vocabulary they will use when interacting with the system. Noyes (1993) notes that some manufacturers of speech recognition technology claim that their system requires only one pass of enrolment for each word in the vocabulary. Baber (1991a) counters this claim by pointing out that one utterance of each word would not be able to provide enough information on speech variation for the system. He goes on to say that the optimal trade-off between content and time involves each word in the vocabulary being repeated between 3 and 5 times for enrolment. In contrast, for speaker-independent systems the vocabulary accepted by the system has been pre-defined from a large number of speech samples from a large number of people in advance of the individual using the system. Baber (1991a) notes that between 1983 and 1989 the share of the market for speaker-independent systems fell from 77% to 60%. Recent trends have continued in the shift towards speaker-dependent systems. This trend looks set to continue as there are still technological problems which have to be overcome before speaker-independent systems achieve an acceptable level of recognition performance.

Although these technological limitations will continue to be addressed in the future there are usability issues which still have to be considered for both speech and keypad input in relation to automated telephone services.

1.3. Know the User

Shneiderman (1987) suggested a range of performance measures as a suitable way of quantifying the efficiency and usability of an interactive system. These are:

- speed of performance
- rate of errors
- subjective satisfaction
- time to learn
- retention over time

These measures have been adopted to a greater or lesser extent by the majority of researchers in the area of user interface design for automated telephone services. However, the results of research reported in the literature indicate that observing the time taken by a user to complete a task, or the number of errors made by a user when completing a task, will not provide the designer with all the information needed to develop a more user-friendly interface. For example, Murray (1991) states that it is widely recognised that some people are more successful at using speech technology than others ('sheep' v 'goats'). However, the factors which lead to this division have not been satisfactorily explained. In addition, Fraser (1995) states that dialogue systems engineering is a young discipline which suffers from the attitude adopted by many software engineers of "implement first, design and specify later".

Interacting with a computer system is a complex cognitive task where performance differences, based on measurements like Schneiderman's, need to be explained instead of just being identified and used as evidence to support a design decision. By explaining the differences in performance between groups of individuals, the dialogue engineer will have information which can be used to improve the usability of the automated telephone services being designed. This thesis therefore examines the proposition that individual differences in cognitive abilities affect performance on computer tasks, specifically in the context of automated telephone services. The role of individual differences in human-computer interaction has previously tended to focus on three areas: text editing, information retrieval and computer programming (Egan, 1988). As previous research into the area of spoken dialogues systems discussed above has shown, no real attempt has yet been made to understand users' behavioural data in order to improve the usability of these type of systems. This thesis presents a detailed investigation of user behaviour relevant the engineering design of telephone-based user interfaces.

In their seminal work, Egan and Gomez (1985) suggest a three step approach to dealing with individual differences in human-computer interaction. Step 1 is the "***assay***" which requires the designer to assess the extent to which individual differences will affect task performance. This involves identifying individual differences through the use of appropriate psychometric tests. For example, if the designer is interested in the role of personality, subjects complete a personality test to allow the designer to compare performance with personality characteristics in order to identify if certain personality types perform better than others. Step 2 requires the designer to "***isolate***" where the salient individual differences identified in the first step have their greatest effect in the interaction such as which sub-tasks account for the greatest variance in performance. In step 3, the designer "***accommodates***" the important features which have been isolated in step two by including appropriate features and measures in the user interface design. This may involve changes to the interface design or changes in the written instructions given to the user on how to use the system.

1.4. Structure of Thesis

The research work reported in this thesis applies this three step approach to user interaction with spoken dialogue telephone systems. Chapter Two provides an introduction to psychological studies of individual differences with particular emphasis on previous research carried out on the study of individual differences in human-computer interaction. Chapter Three provides the methodological framework on which the experimental work of the thesis was based. This includes a detailed account of the Wizard of Oz simulation technique and a detailed presentation of the Egan and Gomez approach with specific reference to the needs of automated telephone services. Chapters Four, Five, and Six contain experimental results for each of the steps of the Egan and Gomez approach - ***assay, isolate*** and ***accommodate***. Finally, Chapter Seven draws together the results obtained from the experimental work and concludes that, by examining the role of individual differences in spoken dialogue systems for automated telephone services and by attempting to accommodate such individual differences in the design of these types of interfaces, engineers have the opportunity of designing interfaces which are more highly usable by a wider range of users.

Chapter 2: Individual Differences in Human-Computer Interaction

2.1 Introduction

This chapter provides relevant background on the psychology of individual differences. In addition, it presents a review of previous research which has investigated individual differences in human-computer interaction. From the review, it emerges that individual differences play a part in determining whether or not users can perform specified tasks successfully. On the basis of these findings and the results of the research carried out on spoken dialogue systems reported in this thesis, it is shown that there is a need for greater understanding of the differences that exist between individual users and groups of users who use spoken language dialogue systems. By studying the role of individual differences in subject performance, dialogue engineers could be equipped with information on how to improve the usability of automated telephone services and therefore open up these services to greater numbers of users.

2.2. Individual Differences: Historical Background

The study of individual differences dates back to the work carried out by Sir Francis Galton (1869) and Alfred Binet (1905). In 1869 Galton published a book entitled "*Hereditary Genius: An Inquiry into its Laws and Consequences*" which studied the family trees of eminent men and found that there was a strong tendency for eminence to run in families. Galton therefore came to the conclusion that intellectual ability was largely determined by heredity. A fundamental problem with Galton's approach was his almost total disregard of the effects of environmental influences. The approach adopted by Galton can be attributed to social attitudes and beliefs which were held by the privileged classes in his time. Having said this, Galton's contribution to the study of individual differences lies in the fact that he was probably the first person to have studied them systematically and he formulated many of the key concepts and methods which are still used in this area of research. For example, Galton was the first person to take the mathematical concept of the normal distribution and apply it to psychological characteristics. Galton worked on the assumption that as various physical characteristics such as height seem to be to be distributed in this way, psychological attributes, which he believed to have a physical basis, should follow the same pattern. On this point Galton appears to have been correct: many psychological traits do seem to follow a normal distribution, irrespective of their relationship to physical processes.

Galton was interested in the idea of mental testing to support his methodological ideas but never progressed beyond a rather crude level. The first formal mental tests did not appear until 1905 as a result of work carried out by Binet and Simon in Paris.

Binet had been asked by the education authorities in Paris to devise a method to identify those children who were not succeeding in school and who would require special education. He adopted a pragmatic approach and brought together a number of everyday tasks such as counting money, which were supposed to be representative of tasks that required the use of higher mental functions such as comprehension, judgement, reasoning and adaptation. The tasks were arranged in order of difficulty and administered by trained examiners. Binet assigned an age level to each of the individual tasks to mark the youngest age at which a child of normal intelligence could be expected to complete the task. The aim was to start the child being tested with the tasks for the youngest children and progress through the tasks in order of difficulty until they could complete no more tasks. Binet associated a child's 'mental age' with the last tasks that the child could successfully perform. Later, other investigators put forward the suggestion of the ratio of mental age to chronological age, to be used as the basis of an intelligence quotient (IQ). If a child achieved a mental age 'score' of 100 months on Binet's test and his or her actual age was 90 months, their IQ would be recorded as 111.

One objection to this idea of mental age is that Binet appears to have been using different tasks to measure different aspects of intellectual activity at different ages and it is therefore questionable if this can really be regarded as testing a general concept of mental age. Binet himself admitted that intelligence was too complex a concept to be interpreted and explained by a single number and that this number should not be regarded as a label. This is an interesting point to note because over the years Binet's observations on the limitations of intelligence tests have continued to be overlooked or forgotten by psychologists and society in general.

In 1916 Lewis Terman of Stanford University carried out an extensive revision of the Binet test. This became known as the Stanford-Binet test and it covered not only children but also adults. The Stanford-Binet test set the standard for many of the written mental ability tests which were to follow.

The work of Charles Spearman represents the next major step in the study of individual differences. Spearman's main contributions to this area lie firstly in his work on factor analysis which has had a significant effect on the whole field of mental testing since the 1930s; and secondly in his two-factor theory of intelligence.

In the early 1900s Spearman was administering a variety of mental ability tests to children. Adopting the new statistical technique - correlation coefficients - he correlated the resulting test scores allowing him to investigate relationships between the various sub-test scores. Spearman found that the scores were positively correlated and a child who had a high score on one test also tended to score highly on another. These results led Spearman to conclude that all the tests had something in common and he labelled this factor *g* for 'general intelligence'. In fact Spearman held the strong belief that the factor *g* entered into all intellectual tasks. Spearman also found that there were 'specific elements' or abilities, which he labelled as *s*, that were not correlated with each other. As a result, he regarded each individual test as having an overall *g* loading and a specific loading, *s*. The theory of intelligence that Spearman developed was therefore composed of two factors where general intelligence, *g*, affects all mental abilities, whereas specific abilities, *s*, are not significantly correlated with each other.

Spearman regarded intelligence as being composed principally of *g*, with an influence on every area of mental life. Individuals differ in the amount of *g* they possess - 'bright people' were considered to have high *g* whereas 'dull people' were considered to have low *g*. In addtion, *g* was seen as explaining why people who were good at one mental activity tended to be good at others. People do however differ according to their specific *s* abilities and therefore one person may be better at composing letters than fixing a car, regardless of the fact that these activities are under the overall influence of *g* and are therefore correlated.

Implicitly linked to this theory of intelligence was Spearman's development of the statistical technique known as factor analysis. Factor analysis is a statistical technique which aims to find factors (or 'hypothetical constructs') which explain the relationship between a subject's scores on several tests or sub-tests. As Coolican (1994) shows, there are several steps involved in this statistical procedure.

- A large sample of people need to be measured on several tests or subtests.

- Correlations are calculated between every possible pair of tests or subtests and arranged in a matrix.

- The matrix of correlations is fed into the factor analysis program which looks for 'clusters' - groups of tests or subtests which all correlate well together.

- The researcher sets the program to solve the matrix for a particular number of 'factors'. Factors at this point are mathematical concepts which will 'account for' as much as possible of the correlations found. The program then gives the best configuration of this number of factors to account for all the correlations.

- Alternatively, the program will offer a solution in the best number of factors, with the least amount of variation unaccounted for. The whole 'explanation' is purely statistical accounting for the numerical relationships.

- The researcher might use the program to solve for a higher number of factors if the amount 'unexplained' is too high.

Coolican (1994) also provides a clear tutorial example of the concept of factor analysis concerning several hundred people, of average fitness performing various athletic events. Their performance on every event is correlated with every other to produce a correlation matrix, as shown in table 2.1.

	100m	200m	3000m	5000m	shot	discus	long jump
100m		0.87	0.24	0.31	-0.65	-0.32	0.47
200m			0.19	0.28	-0.61	-0.29	0.39
3000m				0.91	-0.16	0.03	0.13
5000m					-0.08	0.11	0.09
shot						0.65	0.14
discus							-0.02
long jump							

Table 2.1. Correlations between various athletic events

The example shows a strong correlation between performance in the 100m and 200m, and then between 3000m and 5000m events. There is also a moderate correlation between discus and shot put and between 100m and the long jump. The factors underlying the relationships between these results could be named intuitively as "sprinting ability", "stamina" and "strength". If the factor analysis program was used to solve for just two factors based on these results, it would probably show that, regardless of the way the correlation matrix was solved, much of the relationship between the variables was left unaccounted for. If three factors were chosen instead, there would probably be a good solution offered with a small amount of variation left unexplained. However, it should be noted that it would be left to the experimenter to name the factors and explain what processes they indicate.

This same method of factor analysis has been used for large samples of subjects on personality and intelligence tests and the factors which have emerged have been recognised and named, intuitively, by the researchers. These factors are said to be responsible for subjects' variation in performance across the various tests and subtests. Whilst factor analysis does not prove the existence of these factors, it does offer the researcher indicators that personality or intelligence could be organised in a particular way. Factor analysis is only a statistical process.

Having discussed the early developments in the study of individual differences, attention will now be focused on those individual characteristics which may be relevant, in terms of individual performance, to human-computer interaction. These include personality (considered to the most stable of all individual characteristics), memory, learning and other cognitive skills such as verbal ability and spatial ability.

2.3. Individual Differences in Personality

Definition of the concept of personality is essentially concerned with the task of defining the consistency and continuity in the qualities which individuals display. Atkinson, Atkinson and Hilgard (1983) provide a good working definition of personality as follows: "Personality describes the characteristic patterns of behaviour and modes of thinking that determine an individual's adjustment to the environment".

The idea that individuals can be classified into distinct types goes back to Hippocrates (c.400BC) who argued that individuals could be regarded as falling into one of four categories: choleric (irritable), melancholic (depressed), sanguine (optimistic) and phlegmatic (calm).

More recently Carl Jung (1971) argued that people fall into two major categories. One group he labelled as ***introverts*** who are regarded as being shy, preferring to be alone, indulging in solitary activities rather than seeking social interaction. The other group are labelled ***extroverts*** and are seen as actively seeking the company of others and enjoying participating in group activities.

In these typologies the groups are regarded as being distinct and non-overlapping. An alternative approach to the study of personality assumes that individuals differ on a range of continuous dimensions or 'traits' where the differences between individuals are regarded as quantitative rather than categorical and qualitative. On the whole, the type approach has generally given way in personality psychology to the trait approach, as the assumption that people fall into discontinuous categories has come to seem untenable to most observers (Carver and Scheier, 1992).

2.3.1. Personality Trait Systems

The fundamental issues facing trait-based approaches to the study of personality concern the classification and number of basic traits which can be used to effectively describe the basic structure of personality. In the quest for answers to these questions trait psychologists have been helped a great deal by factor analysis. Two of the most important contributors to the development of the trait approach have been Raymond Cattell and Hans Eysenck.

Cattell (1946) believed that the traits which are basic to personality could only be determined empirically and his research involved the use of empirical methods to determine the structure of personality. As a basis for his work Cattell took a set of 4,500 words, each of which gave a description of personality based on a much larger set created by Allport and Oddbert (1936) and from this set removed synonyms and metaphorical terms. This process left a total of 171 trait descriptions for which he collected ratings from factor analysis. The factors which were produced as a result of this analysis indicated those trait dimensions which he believed were important in describing human personality. Cattell's work also emphasised the importance of a multivariate approach to the study of personality involving the collection of self-report questionnaire data, observer ratings and objective behavioural data. Measurement of the structure of personality is a complex issue which requires a fully rounded research method. Using data from thousands of subjects the study concluded that the basic structure of personality consisted of 16 dimensions (primary factors) which provided the name for the inventory that was designed to measure them: Sixteen Personality Factor Questionnaire (16PF) (Cattell, Eber & Tasuoka, 1970). Each of the 16 factors is represented by separate ratings and self report responses. An example of the kind of statement used in this personality inventory is:

"I can forget my worries and responsibilities whenever I need to"

Subjects choose one of three answers: a: yes, b: sometimes, c: no. Each subject obtains a score which places them at some point along each trait dimension for each of the 16 factors. Table 2.2 gives an example of the personality dimensions measured by this inventory.

COOL	Versus	WARM
CONCRETE-THINKING	Versus	ABSTRACT-THINKING
AFFECTED BY FEELINGS	Versus	EMOTIONALLY STABLE
SUBMISSIVE	Versus	DOMINANT

Table 2.2 Example from Cattell's 16PF test

Eysenck's (1975) starting point in the search for the basic dimensions of personality structure was somewhat different from Cattell's since he believed that the investigation should start with *a priori* definitions of the traits to be measured and how they might be reliably and validly measured. This can be contrasted with Cattell's "lexical criteria of importance" approach which is based on the idea that "a quality of personality that is described by lots of words is likely to be of more importance than one described by just a few" (Carver and Scheier, 1992, p66). Eysenck's work investigated the typology of Hippocrates and related observations made by Jung (1971) (people fall into one of two categories, introverts and extroverts) with the hypothesis that the four basic types listed by Hippocrates and Jung could be created by merging the high and low levels into what he termed two "super traits". These two super trait dimensions, which characterise the basic structure of personality were labelled *Introversion-Extroversion* and *Emotionality-Stability* (previously labelled as neuroticism). An extrovert is characterised as someone who tends to be sociable, lively, active, dominant whilst an emotional individual is characterised as being someone who easily or frequently becomes upset. The observations of Hippocrates and others were recast into a matrix of these two underlying dimensions.

Table 2.3 gives an example of the characteristics of four groups of people who have various combinations of factors in these two dimensions. Classes in italics refer to Hippocrate's typology. As can be seen from Table 2.3, people who are introverted and emotionally stable tend to be controlled and calm in their actions. On the other hand, individuals who are introverted and emotionally unstable, tend to be rigid and reserved in their behaviour. In terms of extroversion, when this is combined with emotional stability the result is usually a carefree, sociable individual. When extroversion is combined with emotional instability the result is an individual who tends to display a restless, aggressive quality in his or her behaviour. The essential feature that distinguishes Eysenck's trait approach from the type approach it subsumes is that the behavioural effect of one dimension (e.g. introversion-extroversion) can be different depending on what other traits the individual may have (e.g. emotional instability).

	Emotionally Stable		**Emotionally Unstable**	
	passive		quiet	
Introvert	peaceful		sober	
	controlled	*Phlegmatic*	rigid	*Melancholic*
	calm		reserved	
	sociable		active	
Extravert	lively	*Sanguine*	aggressive	*Choleric*
	carefree		restless	
	leader like		touchy	

Table 2.3 Personality Characteristics according to Eysenck's Trait model (taken from Schier 1992, p69). Italicised words refer to Hyppocrates' typology.

These dimensions can be assessed by the Eysenck Personality Questionnaire (EPQ) (Eysenck, 1975) a self-report questionnaire developed using the factor analysis method. Unlike Cattell, Eysenck did not use the factor analysis method to establish which dimensions of personality existed but instead adopted this approach to improve the reliability of the measurement.

Another approach to the study of personality, the Myres-Briggs Type Indicator (MBTI) (Myres & McCaulley, 1985), has also been developed along similar lines and has been used in several applied areas of personality psychology such as organisational psychology and human-computer interaction.

In the 1950's Myres and Briggs adopted the Jungian personality typing, modified it by adding a fourth scale, simplified its description and developed a psychometric test called the Myres-Briggs Type Indicator (MBTI). This approach to personality typing assumes that an individual's personality can be defined by dividing it into four orthogonal (i.e. independent) preference scales: *energising, attending, deciding* and *living*. Within each scale people have a preference for one of two opposites which define the scale. This results in a total of 16 different combinations, each of which provides a description of a unique personality type. This test has been widely used in business, education, career counselling and research in the area of human-computer interaction.

For each of the four preference scales which underlie this approach to personality typing, each individual is regarded as having a preference for one of the two opposite choices (which are designated by a letter):

- **Energising** - How a person is energised:
  Extroversion (E) - preference for drawing 'energy' from people, activities
  Introversion (I) - preference for drawing 'energy' internally from ideas, emotions

- **Attending** - What a person pays attention to:
  Sensing (S) - preference for using senses to notice what is real
  Intuition (N) - preference for using imagination to envision what is possible

- **Deciding** - How a person decides:

  Thinking (T) - preference for organising and structuring information logically

  Feeling (F) - preference for organising information in a value-oriented way

- **Living** - Lifestyle a person prefers:

  Judgement (J) - preference for living a planned and organised life

  Perception (P) - preference for living a spontaneous and flexible life

The commonly accepted order for describing each of the combinations is energising: attending: deciding: living. For each of the personality types detailed, personality profiles have been assembled from many years of application and analysis on large populations of subjects. The result (claim the advocates of this approach) is that these personality profiles give a description of the personality of most adultpeople. This according to Noring (1993) is backed up by corroboratinganecdotal evidence expressed by individuals after reading their personality profile.

2.3.2. A Comparison of Trait Methods

What is interesting to note in comparing the work of Cattell and Eysenck are the similarities in personality structures produced by both methods. The two dimensions which Eysenck identifies as being the super-ordinate dimensions of personality are similar to the first of the 16 factors listed by Cattell (Cool - Warm, Affected by Feelings - Emotionally Stable). This similarity becomes even stronger when *second order factors* from Cattell's scale are considered. Second order factor analysis identifies which factors themselves actually form factors and are correlated in clusters.

As a result of this type of analysis Cattell acknowledged that one second order factor deriving from the 16PF scale is very similar to Eysenck's *introversion-extroversion* dimension and another, which Cattell labels *anxiety*, is similar to Eysenck's concept of *emotional stability*. Cattell (Cattell and Kline, 1977) has criticised Eysenck for not covering what he regards to be the full range of personality dimensions in the ratings he uses to develop his factors. Eysenck on the other hand argues (Eysenck, 1975) that the two super traits are not the only dimensions of personality and they can in fact be broken down as part of a hierarchy of qualities which combine to make up personality. For example, the dimension known as 'extroversion' is composed of several traits which themselves are made up of habitual responses, which in turn are derived from a sets of specific responses. Another point which should be noted about Eysenck's analysis of personality is the idea that the two dimensions are related to certain aspects of nervous system functioning. For example, Eysenck believes that certain processes in the brain underlie the dimension of introversion-extroversion. Eysenck argues that extroverts are more cortically aroused than introverts which results in the extrovert seeking stimulation with the introvert trying to avoid over-stimulation. In addition, Eysenck's analysis of personality highlights a third dimension (psychoticism), though not to the same extent as the first two. This dimension refers to the individual's predisposition toward becoming either psychotic or sociopathic (a personality disorder characterised by maladaptive social relationships that reflect anti-social behaviour). Eysenck regards people high in psychoticism to be cruel, hostile and unconcerned for the welfare of others.

2.3.3. Towards a Consensus on the Basic Dimensions of Personality

As has been shown, there is a wide diversity in how people have approached the problem of identifying the basic traits to personality structure. What does seem to be emerging in the field of trait psychology however is a degree of consensus on the idea that the basic structure of personality may only consist of five super-ordinate factors, which Goldberg (1981) has labelled as the "Big Five". The evidence which supports a five factor theory of personality is not new - rather it has evolved over a period of 40 years. Digman (1990) provides a detailed account of this process. The first evidence supporting the five factor model was found by Fiske (1949) who reported from his own analyses that he was unable to replicate the 16 factor structure claimed by Cattell. Instead, he claimed to have found a five factor solution. These findings seemed to have gone largely unnoticed until the 1960's when there was a renewed interest in this issue. Norman (1963), Borgotta (1964) and Smith (1967), independently reached the same conclusion: that the data was best accounted for by adopting a five factor approach. In the 1980's there was a further burst of research activity aimed at establishing the existence of the "five factors" basic to personality. Data were collected from different and more diverse samples than before. For example Digman & Inouye (1986) collected data based on teachers' ratings of children. McCrae & Costa (1987) used self-report and peer ratings as a source of data. Botwin and Buss (1989) measured the frequency with which individuals indulge in certain types of actions (these data were collected from both self-reports and observer ratings).

Although there has been a move towards a consensus view that a five factor model adequately describes the basic characteristics of personality, no consensus has developed on how these five dimensions should be labelled. This can be attributed, in part, to the intuitive naming process required by factor analysis which may reflect an underlying bias on the part of the researcher. As a result of this the five factors have been given a variety of names, as Table 2.4 shows.

Factor	Norman (1963)	McCrae & Costa (1987)
1	Surgency	Extroversion
2	Agreeableness	Agreeableness
3	Conscientiousness	Conscientiousness
4	Emotionality	Neuroticism
5	Culture	Openness to Experience

Table 2.4 Example of names used to label the "Big Five" personality traits

Factor 1 is characterised by the assertiveness of the individual, their sociability. Factor 2 incorporates the idea that the individual is very emotionally supportive of others, whereas the individual who scores low is seen as having antagonistic and oppositional qualities. There is some disagreement about the nature of the third factor, though most theorists label it as conscientiousness, which implies purposefulness and the striving towards goals. The fourth factor is regarded by most as representing what Eysenck had emphasised: emotionality or neuroticism. The fifth factor also causes some disagreement amongst the theorists. Costa & McCrae (1987) have labelled this as openness to experience.

2.3.4. The Comprehensiveness of the Five Factor Model

The degree of comprehensiveness of the five factor model can be considered by comparing this model to previous Cattell and Eysenck models of personality structure. In comparison with Eysenck's model, two of the five factors (Factor 1 and Factor 4) represented are similar to Eysenck's super-traits of *extroversion* and *emotional stability*. Even Eysenck's third dimension is covered by Factor 3, which has been labelled as *conscientiousness* by other five-factor theorists. Another aspect of Eysenck's work that seems to have been adopted by the five factor theorists is the idea of the five factors as super traits which include subordinated traits within them. For example, Costa & McCrae's NEO Personality Inventory - Revised (NEO PI-R) (Costa & McCrae, 1992) has six measures for each of the five super traits.

In comparison to Cattell's system the most obvious difference is in the number of traits identified : five as opposed to sixteen. No other researchers have been able to find more than seven factors as making up the basic structure of personality (Goldberg, 1981, Digman,1990). In Cattell's favour however is the fact that when his sixteen factors were subjected to second order factor analysis, the two dimensions which emerged were similar to Eysenck's two super traits (Cattell & Kline, 1977).

The five factor model can also be compared to Wiggins' (Wiggins, 1979) concept of the interpersonal circle. Again similarities can be seen between the two models. For example, the basic dimensions of the interpersonal circle are *dominance* and *love. Love* can be seen as being very similar to Factor 2 in the five factor model which has been labelled as *agreeableness.* If *dominance* is taken to be similar to Factor 1 in the five factor model which has been labelled *extroversion,* it can be said that the interpersonal circle is made up of the first two factors in the five factor model.

Finally it is possible to compare the five factor model with the Myres-Briggs Type Indicator. The first factor that the MBTI measures is *extroversion,* which is covered by the first factor in the "Big Five" model. The second factor in the MBTI scale has been labelled as *Intuition,* this can be said to be similar to the fifth factor in the Big Five model which has been labelled as *intellect/openness to experience.* The third factor in the MBTI scale is *feeling* and this can be regarded as being similar to factor 2 in the Big Five model which has been labelled *agreeableness.* The final factor in the

MBTI scale has been called *judgement.* When this dimension has been compared to the five factor model it has been identified with factor three in that model: *conscientiousness.* The five factor approach is therefore similar to the MBTI scale. The difference is that the five factor model offers a more comprehensive analysis of the basic structure of personality as can be seen by the inclusion of a fifth dimension not covered by the MBTI scale; namely factor four: *neuroticism.*

To reflect the move towards the Big Five consensus in the study of personality structure the NEO-PI-R (Costa & McCrae, 1992) personality inventory was the method adopted to assess personality in the experiments reported in this thesis. It was chosen because it offers a concise measure of the five major personality dimensions and provides information on some of the more important traits that help define each dimension. The NEO-PI-R is based on the idea that personality traits are arranged in hierarchies from very broad to very narrow traits and that these general traits (domains) and specific traits (facets) should be assessed.

Most of the research leading to the development of the NEO-PI-R was conducted on the basis of large scale samples. For example, the *Augmented Baltimore Longitudinal Study of Aging* (Costa et al, 1986) introduced the NEO Inventory, a three factor (Neuroticism, Extraversion and Openness) precursor the the NEO-PI-R into a longitudinal study of aging being carried out by the National Insitute of Aging (Shock et al, 1984), approximately 400 men and 300 women took part in personality assessment as part of this six year study. The sample of subjects consisted largely of individuals working in (or retired from) professional or managerial occupations and were considered to be better educated than the population in general. As part of the research which led to the development of the facet scales for Agreeableness and Conscientiousness dimensions of the NEO-PI-R over 1,800 men and women employed by a large national organisation were assessed. This is referred to as the *The Employment Sample* (Costa, McCrae and Dye, 1991). This sample differed from the Baltimore study on which the NEO-PI-R was orginally developed in several ways, providing a more representative sample of the general adult population:

- most participants in the study were highschool graduates, but a significant proportion of them did not have higher education qualifications.

- the participants were considerably younger in age with several hundred of them being aged between 20 -29.

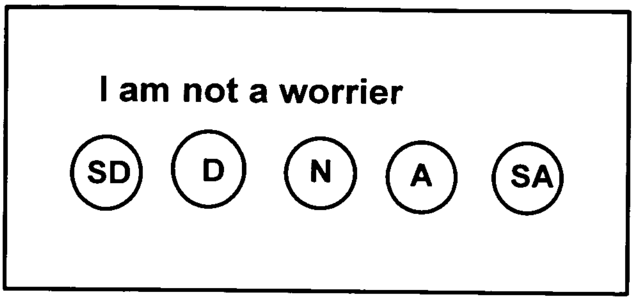
- the sample contained a higher proportion of non-whites - 21% black, 10% Hispanic, Asian or other ethnic minority, than is representative of the United States population.

At each stage of its development the NEO-PI-R shceme, both factor analysis and a rational approach to scale construction was adopted. In terms of scale construction, Costa and McCrae identified the construct they wished to measure and wrote proposing statements that, if they were answered in the key direction, would suggest the presence of the underlying trait. This approach is based on a theory of item responding known as *self-disclosure* (Johnson, 1981) which assumes that individuals will respond candidly when presented with item statements in this manner. In addition, Costa and McCrae were influenced in adopting the rational scale approach by their experience of item factor analysis which almost always produced rationally interpretable factors and indicated that subjects tend to respond to the manifest content of item statements. Item factor analysis identifies clusters of items that covary with each other but are relatively independent of other clusters of items. Factor analysis was also used to identify the main domains and, within domains, to identify facets. This development process produced the five domains and thirty facets which now make up the NEO-PI-R as Table 2.5 shows:

Neoroticism	**Extraversion**	**Openness**	**Agreeableness**	**Conscientiousness**
Anxiety	Warmth	Fantasy	Trust	Competence
Angry Hostility	Gregariousness	Aesthetics	Straightforwardness	Order
Depression	Assertiveness	Feelings	Altruism	Dutifulness
Self-Consciousness	Activity	Actions	Compliance	Achievement Striving
Impulsiveness	Excitement-Seeking	Ideas	Modesty	Self-Discipline
Vulnerability	Positive Emotions	Values	Tender-Mindedness	Deliberation

Table 2.5. Domains and Facets of the NEO-PI-R

Two versions of this inventory are used in the course of the work reported in this thesis. In the first experiment (see Chapter Four) the NEO-PI-R was used. This is a self-administered test, appropriate for men and women of all ages, which consists of 240 questions to be answered by subjects on a five point rating scale. An example of the format is given in Figure 2.1.



I am not a worrier

SD D N A SA

Figure 2.1 Example of NEO-PI-R Format

A subject has five possible responses to give to each statement in the inventory:

- a subject selects *SD* if they strongly disagree with the statement or feel that the statement is completely false.

- a subject selects *D* if they disagree with the statement or feel that it is mostly false.

- a subject selects *N* if they were neutral on the statement or cannot decide.

- a subject selects *A* if they agree with the statement.

- a subject selects *SA* if strongly agree with the statement.

The NEO-PI-R should typically take about 45 minutes to complete but older subjects and those who have limited reading ability may take longer to complete it. This was found to be the case in the first experiment. The NEO-PI-R provides global information on five major dimensions of personality as well as providing scores on each of the six facets which make up each dimension. As the aim of the experiment was to look for a relationship between the five major dimensions and performance on spoken language dialogue systems, only the global scores for the five major dimensions of personality were used in the analysis carried out in Chapter Four. In the second experiment (see Chapter Five) the NEO Five-Factor Inventory (NEO-FFI) which has a similar format to the NEO-PI-R was used. This is a reduced 60-item version of the NEO-PI-R that is scored for the five major domains of personality only. Like the version of the NEO-PI-R used in this first experiment, this inventory was self-administered with the subjects answering each statement using a five point rating scale. The NEO-FFI is considered appropriate for use when the time available to test the subject is limited and the global information on personality provided by this measure is therefore considered to be sufficient for the research purposes of this work.

2.4. Individual Differences in Cognition

The term cognition covers the mental processes of perception, memory and information processing which allow an individual to acquire knowledge, solve problems and plan for the future. Human cognition is the result of the interaction between an individual's internal thoughts and external environment when faced with a task. A task involves the use, storage and acquisition of knowledge. There have been attempts to collate a range of cognitive abilities into "general intelligence" to give an intelligence quotient (IQ), as mentioned earlier. Other research in areas of cognitive activity has focused on learning or second language acquisition.

Although there is no single agreed taxonomy of cognitive abilities, broad differences have been observed repeatedly. For example, Cooper and Mumaw (1985) reviewed the research on spatial ability and concluded that a spatial factor existed which was independent of verbal and quantitative factors. In addition, evidence exists for both holistic learning and serialistic learning styles typical of verbal-learning scenarios where the subject is presented with a list of words in a fixed order and they attempt to recall the words in that order (e.g. Pask, 1976, Entwhistle, 1978, Pask, 1980).

Cognitive abilities are assessed or measured by asking the subject to perform some kind of cognitive task which often takes the form of a paper and pencil test requiring the subject to complete a sequence of numbers, select synonyms for words or choose a diagram from a group of three or four which matches some criterion. Subjects are expected to complete the tasks to the best of their abilities and are usually assessed by how quickly or accurately they can perform the task.

2.4.1. Spatial Ability

Spatial ability refers to the extent to which individuals can deal with spatial relations and the visualisation of spatial tasks. Real-world tasks, such as learning to move through the physical environment, working out the trajectory of an approaching object, as well as more intellectual tasks such as solving complex problems in engineering, seem intuitively to require some degree of spatial skill.

Psychometric work on spatial aptitude has relied on the use of factor analysis. Factors labelled as spatial ability appeared in early factor analysis work (McFarlane, 1925, Thurstone, 1938) and were independent of abilities such as verbal reasoning. Over the last 60 years, a number of tests have been developed to assess spatial ability. An example of one of these tests is Raven's Progressive Matrices (1958) where the subject undertakes a performance test consisting of 60 designs, known as the 'matrices'. The aim of the test is for the subject to select the missing part of the design from a set of alternatives. The designs progress from being simple shapes to designs consisting of abstract logical relations.

Spatial aptitude tests are regarded as being reliable predictors of performance in both scholastic and industrial settings. McGee (1979) carried out a review of research based on the predictive ability of spatial aptitude measures and concluded that for academic and vocational-technical training programmes spatial ability correlated most significantly with course grades in mechanical drawing, art, mechanics and mathematics. In terms of job performance in industry, where productivity and supervisor ratings are used as criterion measures, spatial aptitude tests have been used to predict success in engineering, design and other mechanically-oriented areas. A key finding of McGee's review was that spatial ability tests are able to measure a component in real-world skills not measured by other types of aptitude tests.

With reference to computer-based tasks, a measure of spatial ability can be correlated with an individual's performance with a particular system, the hypothesis being that performance is dependent, in part, on spatial ability. Examples of how this has been applied to human-computer interaction research are provided later in this Chapter.

Spatial ability has a limited role to play in examining performance differences in people's use of automated speech systems. As Martin, Williges and Williges (1990) noted, telephone interfaces require serial presentation of information and as a result it is difficult to include spatial cues about movement through a database. From the information they receive over the telephone, users of an automated telephone service have to determine where they are in the dialogue structure at that precise moment in order to provide the appropriate commands that will allow them to continue and complete their task. A short response delay is also a requirement, since the user has a limited amount of time to respond to a system prompt before there is a "time out" of the speech recognition processor. From a review of previous research discussed in Chapter One, it emerged that designers have tried to overcome the problem of spatial orientation by focusing on issues such as the vocal menu structure and the number of levels and stages that users must progress in the dialogue in order to reach their goal. In addition, the number of choices presented to users in a menu list must be considered. By showing that subjects' differences in performance whilst using an automated telephone service may be due, in part, to individual differences in spatial ability, the opportunity is offered to system designers to take the effects of these differences into account in order to design automated systems that are effective and efficient for a considerably larger number of users.

The spatial ability test used for the research work reported in this thesis is the second part of the AH4 Test of General Intelligence (Heim, 1970). This test is a group test of general intelligence designed for use with a cross-section of the general population. The first part of the test consists of 65 questions which have a verbal bias. The second part of the test is made up of 65 questions which have a spatial bias. The questions in both sections are multiple choice in form. Subjects are given ten minutes to complete as many questions in Part One as they can. The same time constraint applies to Part Two. In each part of the test the questions are arranged in an ascending order of difficulty. The scores from each section can be summed to provide an overall measure of general intelligence.

Figure 2.2 gives an example of the format used for the spatial ability section of the AH4 Test of General Intelligence.

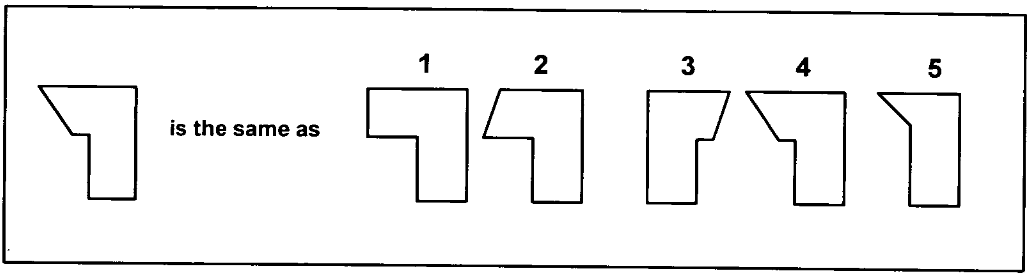


Figure 2.2 Format used for the Spatial Ability Section of the AH4 Test of General Intelligence

The test-retest consistency (reliability) of scores obtained from the AH4 Test of General intelligence was carried out by repeated testing of small groups of subjects, with different intellectual levels, over a 10 week period. The data from these experiments indicated correlations of 0.9 between each test and every other testing thus indicating the reliability of this psychometric measure. Validation of the AH4 can be seen in the norms given below which show the norms rising, predictably, with each criterion.

	Naval Ratings		
	Part I	Part II	Total
Grade A	44-65	53-65	93-130
Grade B	39-43	46-52	83-94
Grade C	30-38	37-45	69-82
Grade D	24-29	31-36	57-67
Grade E	0-23	0-30	0-56
Mean Score = 75.23	Standard Deviation = 14.58		

Table 2.6. Naval Rating Scores for the AH4 Test of General Intellgence

	University Students		
	Part I	Part II	Total
Grade A	59-65	61-65	117-130
Grade B	53-58	56-60	106-116
Grade C	44-52	43-55	89-105
Grade D	38-43	36-42	76-88
Grade E	0-37	0-35	0-75
Mean Score = 96.36	Standard Deviation = 15.01		

Table 2.7. University Student Scores on the AH4 Test of General Intelligence

The range of tasks which rely on a spatial ability component suggests that spatial ability may consist of a number of sub-factors. Theorists disagree on the numbers of sub-factors which comprehensively describe spatial aptitude. Carroll (1983) and Lohman (1979) indicate that there are several reasons for this. Different factor analysis methods have been used which can lead to different interpretations of the factor structure. Also, changes in the administration or format of a test such as the use of time or accuracy as a dependent measure and the complexity of the test items can affect the factor structure which is subsequently derived from the analysis of the test results.

Lohman (1979) tried to resolve some of these inconsistencies by re-analysing most of the major factor-analytic work involving spatial ability carried out in the United States. This resulted in an attempt to redefine spatial aptitude as a general spatial ability factor which had several correlated sub-factors. The following three sub-factors were consistently reproduced by this analyses:

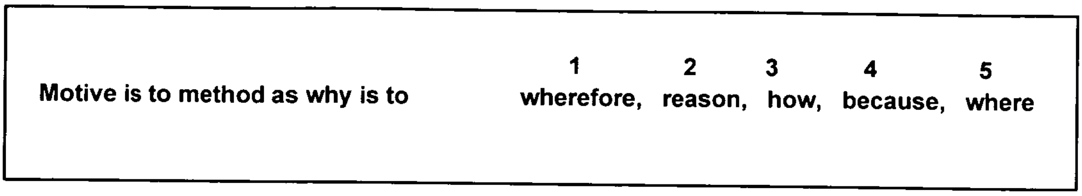
- **Spatial Relations:** the ability to solve simple mental rotation problems quickly.

- **Spatial Orientation:** the ability of the individual to orient himself or herself in space relative to objects and events. This is similar to a factor labelled 'kinesthetic' ('feeling of motion') by Thurstone (1951) and others.

- **Visualisation:** tests that define this spatial sub factor are untimed and usually contain more complex tasks i.e. there is a greater emphasis placed on spatial reasoning.

2.4.2. Verbal Ability

Language is ubiquitous, yet some individuals deal with verbal communications better than others. Verbal ability or 'verbal intelligence' can be objectively assessed by a number of aptitude tests. For example, the verbal ability section of the AH4 Test of General Intelligence requires the subject to answer a set of progressively more difficult questions on the basis of multiple choice. Figure 2.3 gives an example of the type of question subjects are expected to answer.



	1	2	3	4	5
Motive is to method as why is to	**wherefore,**	**reason,**	**how,**	**because,**	**where**

Figure 2.3 Example of Verbal Ability Section Questions from the AH4 Test of General Intelligence

Verbal comprehension is a complex process which is composed of a number of sub-processes including lexical, syntactic-semantic and pragmatic processes.

Lexical Access: This is an automatic (unconscious) low-level process which involves the matching of a string of sounds with stored templates or types. Comprehension cannot proceed unless these tokens have been matched with their concepts.

Syntactic-Semantic Analysis: Once words have been identified, the listener has to derive meaning. Syntactic rules identify the role that each word plays in an expression. This is crucial to meaning. For example, syntax tells us that "John likes Mary" is not the same as "Mary likes John". Semantic analysis enables the individual to decide the meaning of a word in context in an expression by combining the word's denotation with the specific role it has been assigned by a result of the syntactic analysis. "John" in "John likes Mary" and "Mary likes John" has the same denotation but in the first sentence it is the subject; in the second sentence the direct object. The contextual meaning of the word is therefore different because of the different syntactic role of the name.

Pragmatic Sub Process: This requires the individual to interpret the meaning of a message in terms of his or her understanding of what is happening. An example of this can be seen in the meaning that would be derived from the following sentence "John saw the Grand Canyon flying to California". The emphasis here is on context. If John was watching a science fiction film say it might be reasonable to assume that he could see the Grand Canyon flying to California otherwise one would assume that John saw the Grand Canyon when he was in an aeroplane flying to California.

These sub-processes appear to build upon one another. Sentence analysis relies on lexical analysis and in turn, text comprehension depends on sentence comprehension. Although the different components which make up verbal comprehension are distinct, individuals who acquire skills in one sub-process are more likely to display the same level of skills in another. Hunt (1985) sees this as a reconfirmation of the position that verbal comprehension behaves, statistically, as if it were a unitary ability.

Verbal ability has a key role to play in determining subjects' performance with automated telephone systems. How subjects comprehend the information provided by the service affects their behaviour at each stage of their interaction with an automated telephone service. The variance obtained in user error rates for a particular system could be attributed to differences in verbal ability. A subject with low verbal ability may have difficulty comprehending the instructions given by the system to make a choice from a menu and as a result say or enter the wrong response to the system prompt. As system designers move towards the idea of a more conversational approach to spoken dialogue interaction there will be an increase in the importance of verbal skills in determining the individual's perception of the usability of the system.

2.4.3. Cognitive Style

Cognitive style refers to the way in which cognitive tasks are approached or handled. For example, Witkin and Goodenough's theory of Psychological Differentiation (1981) offers a global theory of social, perceptual and cognitive functioning. For cognition, Witkin and Goodenough suggests the idea that there are basically two different types of cognitive style: *field-dependency* and *field-independency,* which reflect different ways of processing and responding to information.

Witkin developed the Embedded Figures Test (1981) as a metric for this theory. This test assesses, through a series of scenes, the individual's susceptibility to contextual cues or their ability to see a figure separate from its background. This test also gives an indication of the individual's approach to learning and problem solving. In the field-independent cognitive style the individual is regarded as having good analytical and reconstructing skills (such as separating an item from an organised field) whereas an individual who is field-dependent is seen as adopting a more holistic approach, with an underlying reliance on the inherent organisation of the material being presented.

This test was not considered appropriate for use in the experimental work reported in this thesis as users of automated telephone services receive information from the service serially and are restricted in the way they can respond to information provided by the service. The value of the Witkin and Goodenough method of cognitive style analysis can perhaps be seen when individuals are interacting with multimedia systems. In this type of system the users are presented with information in the form of different modalities (e.g. speech, text, graphics) and how they respond to this information is dependent on how they focus on the information that is of most relevance to them and the goal they are wanting to achieve.

Another approach to the study of cognitive style differences is the Impulsivity-Reflectivity dimension. The hypothesis here is that when it comes to solving problems some people are "impulsive" and react with the first response they can think of, whereas others are "reflective" and adopt a more systematic and analytic approach.

2.4.4. Learning

When it comes to learning it is important to consider factors which have an influence on the learning process. Of particular interest, especially for issues concerned with the effects of individual differences on performance, are factors such as prior knowledge and learning style.

To learn a new task, adult learners will call upon prior knowledge which is in some way similar to the new situation facing them. For example, if an individual is faced with a new situation which consists of a computerised version of a previously experienced task, such as ordering goods from a catalogue over the phone, the user will evoke knowledge of the rules governing how to do that former task. They will then be faced with aspects of the situation which are new to them and involve having an understanding say, of what the computer can and cannot do. Certain aspects of this new situation such as the information provided by the system, may evoke different kinds of associations for different users and this could lead to some individuals making erroneous assumptions about how the system functions.

This example highlights the role of transfer effects in new situations. Transfer of prior knowledge to new situations can have quantitative and qualitative effects. The qualitative effects relate to the direction of the transfer - does the prior knowledge that the individual has seem to facilitate or hinder new learning?. Quantitative effects relate to the strength of the transfer - to what extent will learning be facilitated or hindered? In the case of the example above, if the individual normally would speak freely over the telephone to an operator who takes their order, they may experience negative transfer effects when their speech is controlled by an automated system with limited speech recognition capability.

Another important aspect of learning concerns the methods of problem solving which individuals adopt when learning a complex task. For example, some individuals are "systematic" whereas others adopt a more "heuristic" approach to problem solving. On the whole, people classified as being systematic are regarded as using abstract logical models or 'schemata', while heuristic people use past experience and intuition. This of course, is dependent on the nature of the task. If a task is well structured it may be better to approach it in a systematic way. Other tasks may be less structured and are consequently difficult to analyse. When faced with a decision-making task in an unstructured environment, the adoption of a more heuristic approach may produce better results. There is also the possibility of a mismatch between the user's learning style and the computer's presentation style which could affect how quickly the user moves from novice to expert.

Users may also be said to adopt an *operation* as opposed to a *comprehension* learning approach. This dichotomy is based on Pask's "Conversation Theory" (Pask, 1976) and was adopted for use in human-computer interaction research by van der Veer (1990) who found individual differences in structuring and storing information to be relevant to tasks which were concerned with the handling of the computer itself such as programming and system design.

It is also possible to distinguish between individuals who are *reflective* and *impulsive* in their behaviour when reacting to certain situations. Waern (1989) regards an individual's predisposition to one of these styles as having a role to play when it comes to learning to use computer systems. A reflective person would reflect more on the problem and would evaluate the other alternatives available more than the impulsive individual, and as a result would make fewer mistakes but would take longer to complete the tasks.

2.4.5. Memory

Any analysis of normal human memory must consider the structure of the memory system in terms of the way in which the memory system is organised and the processes operating within that memory structure in terms of the activities which occur within the memory system.

Several memory theorists (Atkinson & Shiffrin, 1968; Waugh & Norman, 1965) have described the basic structure of memory in terms of a number of stores. Multi-store theorists claim that there are three types of memory store:

- **Sensory stores:** which are modality-specific and hold information briefly
- **Short-term memory store:** which has a limited capacity for information
- **Long-term memory store:** which has unlimited capacity and can hold information for long periods of time.

This has led to the development of a model of memory as shown in Figure 2.4.

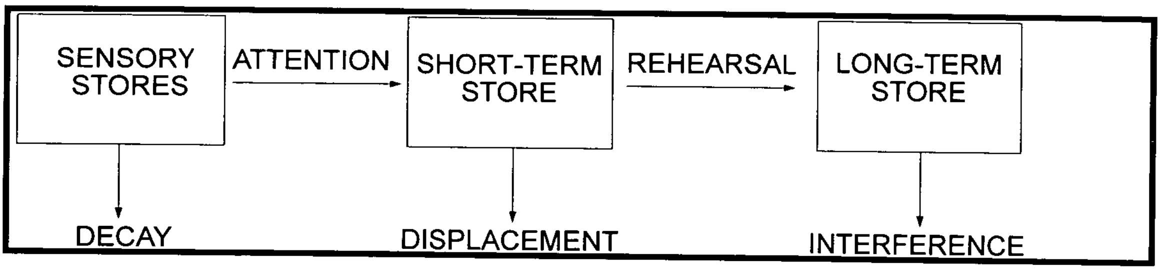


Figure 2.4 Multi-store model of memory (taken from Eysenck, 1990)

This model assumes information comes in from the environment via the sensory stores. These are modality specific where each of the sensory modalities (e.g. hearing, vision) have separate sensory stores. Information is held very briefly in the sensory stores with some of it being consciously processed before it is passed on to be processed further by the short-term store. Information which is not consciously processed will be lost (a process referred to as decay). Information which is held in short-term memory can be lost (displacement) due to the limited capacity of short-term memory for holding information. However, information which is processed in the short-term store can be transferred to the long-term store. This long-term storage of information is often dependent, according to Atkinson & Shiffrin (1968), on rote rehearsal - with a direct relationship between the amount of rehearsal that takes place in the short-term store and the strength of the stored memory. The strength of the stored memory could also be subject to interference. This interference could be *retroactive* whereby the learning of new material, for example, could interefere with the recall of information learned previously in relation to a similar type of task

This model makes a number of implicit structural and processing assumptions. The memory stores form the basic structure and the processes which are labelled *attention* and *rehearsal* control the flow of information between the stores. The emphasis in this model is placed on structure rather than the processes which are operating within that structure.

The multi-store theory was perhaps the first theory of memory which was able to describe the processes and structures of the memory system. The conceptual distinctions it makes are still valid since the memory stores have been shown to differ from each other in temporal duration, storage capacity and forgetting mechanism. This model does however have certain limitations which have to be considered. The major problem is that it is over-simplified. The multi-store approach to memory assumes that the short-term store and the long-term store are unitary (i.e. each store operates in a single, uniform way). Research into short-term store has shown that this is not the case. Warrington & Shallice (1972), carried out investigations with brain-damaged patients and found that one of them had an impaired short-term store but an intact long-term store. They found that the subject's short-term memory deficit affected retrieval of verbal materials like words and digits but did not affect meaningful sounds like a telephone ringing or a dog barking.

The multi-store model also oversimplifies long-term memory. Eysenck (1990) contends that there is a lot of information stored in long-term memory and it seems intuitively wrong to claim that it is all stored in exactly the same way in a single long-term memory store.

The value of the multi-store theory is that it has provided a framework which can be used for examining human memory. The working memory model (Baddeley and Hitch, 1974) was developed for the purpose of trying to eliminate some of the deficiencies in the multi-store model. Baddeley and Hitch (1974) examined the criticisms of the existing theories of short-term memory and argued that short-term memory should be replaced by a new concept which they labelled *working memory.*

Baddeley and Hitch (1974) argued that the working memory system has three components:

- A modality-free central executive which resembles attention in the multi-store model.
- An articulator loop which holds information in a phonological form (i.e. speech-based).
- A visual-spatial sketch pad which specialises in visual and/or spatial coding.

The central executive is the most important component in working memory. It has limited capacity and is used when the individual is dealing with cognitively demanding tasks. The articulatory loop and the visual-spatial sketch pad are the 'slave systems' which are used by the central executive for specific purposes.

Baddeley and Hitch obtained information about the articulatory loop from a word-span study (Baddeley, Thomson and Buchanan, 1975). This study revealed that subjects provided immediate serial recall of approximately the same number of words that they could read aloud in two seconds. This suggested that the capacity of the articulatory loop is determined by time duration in the same way as a tape loop. The characteristics of the visual-spatial sketch pad are not as clear however. Baddeley and Lieberman (1980) distinguished between visual and spatial coding and discovered that spatial coding was more important than visual coding in a variety of tasks leading to the tentative claim that the visual-spatial sketch pad relies primarily on spatial and not visual coding.

This model of working memory has a number of advantages over the multi-store approach to short-term memory. Firstly, this system is concerned with the active processing and temporary storage of information and therefore appears to be relevant to activities such as mental arithmetic (Hitch, 1978), verbal reasoning (Hitch and Baddeley, 1974) and comprehension (Baddeley and Hitch, 1974) and other memory tasks as well. Secondly this model has verbal rehearsal as an optional process which is found in only one of the components of working memory (i.e. articulatory loop). This is in contrast to the significance given to verbal rehearsal in the multi-store model.

Baddeley (1986) later produced a revised version of this model where he distinguished between a phonological store (i.e. speech-based) and an articulatory control process. Baddeley stated that the articulatory loop consists of :

- a passive phonological store which is concerned with speech perception
- an articulatory process which is linked to speech production

Phonological information about words can enter into the phonological store in three different ways - directly through auditory presentation; indirectly through sub-vocal articulation or indirectly via phonological information stored in long-term memory.

This revised model has provided a clear description of the functions of the articulatory loop (i.e. it has a crucial role to play in most short-term memory tasks involving verbal material) but there is still some clarification needed on the precise functioning of the central executive.

Memory has a role to play in determining a subject's performance with a spoken language dialogue system. As Murray, Frankish and Jones (1991) pointed out, the memory loadings imposed by a task and a user interface must be considered by spoken dialogue interface designers. For example, short-term memory limitations may affect performance on a task that requires users to traverse a large number of levels in a menu structure in order to reach their desired goal. In the experiments reported in Chapter Five and Chapter Six of this thesis, subjects have to navigate their way around a menu-driven hierarchically-structured automated music catalogue using touch-tone and speech input modes. Short-term memory limitations may also effect a task where users must remember a large amount of information in order to proceed with their interaction. In addition, if users are frequently required to refer back to previous transactions there is a danger they will not remember what they did or how they got there. The skill of the dialogue engineer is to provide users with an interface that takes into consideration the serial processing involved when using a speech interface in order to reduce the effect of short-term memory limitation as a potential source of user error.

A number of tests have been developed to assess memory ability. For the work reported in this thesis, the Rey-Auditory Verbal Learning Test (AVLT), (Lezak, 1983) was used. This test can be used to assess both long-term and short-term memory. Subjects are given five trials to learn 15 words. They are asked to recall the words after each trial. The subjects are also given a second list of 15 words to recall after the fifth trial, this is known as the 'interference list', after the subjects have recalled all the words they can remember from the interference list they are asked to try and recall as many of the words from the original list as they can. Table 2.8 gives the trial list and interfernce list of words subjects are asked to recall.

Trial List	Interference List
Drum	Desk
Curtain	Ranger
Bell	Bird
Coffee	Shoe
School	Stove
Parent	Mountain
Moon	Glasses
Garden	Towel
Hat	Cloud
Farmer	Boat
Nose	Lamb
Turkey	Gun
Colour	Pencil
House	Church
River	Fish

Table 2.8 Rey Auditory Verbal Learning Test Lists

This test gives an immediate measure of memory span, provides a learning curve, and also measures proactive interefence on memory tasks. Validation of this test can be seen in the norms given in Table 2.9 which show a difference in performance (measured in terms of the mean number of words recalled per trial) between age groups and social class.

Social Group		Trials				
		1	2	3	4	5
Labourers	Mean	7.0	10.5	12.9	13.4	13.9
	SD	2.1	1.9	1.6	2.0	1.2
Professionals	Mean	8.6	11.8	13.4	13.8	14.0
	Sd	1.5	2.0	1.4	1.1	1.0
Students	Mean	8.9	12.7	12.8	13.5	14.5
	SD	1.9	1.8	1.5	1.3	0.7
Old Labourers	Mean	3.7	6.6	8.4	8.7	9.5
	SD	1.4	1.4	2.4	2.3	2.2
Old Professionals	Mean	4.0	7.2	8.5	10.0	10.9
	SD	2.9	2.9	2.5	3.3	2.9

Table 2.9 Adult Norms for Rey Auditory Verbal Learning Test (mean number od words recalled)

2.5. Individual Differences in Human-Computer Interaction

In the field of human-computer interaction, the aspect of individual differences which has received the most attention is the individual's prior knowledge. This would include a person's experience of the task to be performed, the system to be used or their experience of the system as a tool for the task in hand.

2.5.1. Prior Knowledge, Learning Styles

Egan (1988), states that when users who have different experience are compared, priorexperience is often the most powerful predictor of performance. In a study of text editing performance (Rossen, 1983), it was found that experience correlated significantly with the number of lines edited per minute. Elkerton and Williges

(1984) stated that previous experience accounted for more variance in information search times (in a study which was examining information retrieval strategies in a file-search environment) than specific design variables. Egan (1988) also views variation in performance as being related to subjects' previous experience. Although attributing the cause of performance differences, in part, to the different levels of experience, it must also be considered how the individual's differing cognitive abilities and learning strategies affect how the individual progresses from novice to expert status. For example, no mention is made by Elkerton and Williges of how the knowledge of the expert differs from that of the novice in terms of the amount of information that a novice can "chunk" in comparison to an expert.

Designers need to know how the system communicates the 'basics' to the user and whether this is done in the most appropriate way. Van der Veer et al (1985) use the term 'metacommunication' to describe how the system communicates information about the conceptual model which underlies the design of the system. Most of the work carried out in the area of conceptual modelling has focused on knowledge representation. The effects of individual differences on conceptual modelling can be examined in the light of Witkin's concept of field-dependent and field-independent individuals. Field-dependent individuals develop their model of the system by adopting a 'hands-on' approach. The presentation and form of general instructions should accommodate this fact. The field-independent user may wish, on the other hand, to have a clear mental model of the system before beginning to interact with the system. This could be achieved by use of instruction manuals or by drawing on the user's experience with other systems.

In trying to develop a system model of the user while also deriving some insight into the user's model of the system, it is also important to consider other cognitive factors such as verbal ability, spatial ability and the effects of personality factors. If salient individual attributes are identified and taken into account (this is dependent on the type of interface the user has to use in order to perform their specific task), this could lead to better user performance (measured for instance by time and the number of errors which the user may make) which could also have an effect on the perceived usability of the system.

2.5.2. Individual Differences in User Interface Performance

To address successfully the issue of the effects of individual differences in the field of spoken user interfaces requires the development of an experimental paradigm which allows investigation of the effects of individual differences on computer-based tasks of this kind. An example of an attempt to address this issue can be seen in the work of Jennings and Benyon (1989) a study which set out to investigate how users' previous experience of computer-based tasks and cognitive ability affected their performance when using a database system with various types of interface. The five types of interface styles used in this study were:

- question & answer
- menu
- command
- icon
- button

Jennings states that in the area of computer system design, there is not enough known about those characteristics of users which may affect their performance and enjoyment when using database systems. This in turn is related to a lack of understanding of the demands that different dialogue styles make on users. Van der Veer (1990) stated that people can adapt, to an extent, to an interface design which might not suit them. Users also, however, have some capabilities and certain personality and cognitive characteristics which cannot be easily changed. The study by Jennings and Benyon considered whether users of database systems would prefer to use interfaces with different dialogue styles which would suit their individual capabilities. For example, some dialogue styles require the user to remember complex syntax, whereas others guide the user and provide all the options for output. Jennings felt that it was important to study dialogue style since there have been many assumptions made about which dialogue styles are suitable for different users of database systems but little actual research has been carried out to test these assumptions. The primary focus of the study was to examine the relationship between long term (relatively stable) user characteristics and the use of database system interfaces after an initial learning phase. This involved assessing how these user characteristics related to the individual's performance on the database system with a number of different dialogue styles. For example, the menu style interface for the database system minimised navigation and constrained the dialogue whereas the dialogue style which was used for the command line interface was more open and flexible.

In these experiments the database supported a single task which involved obtaining lists of items available from a mail order shopping catalogue. Users had to specify the item in which they were interested and query the database to obtain information such as: "***How many types of women's T-shirts are available, which cost less than fifteen pounds, are navy in colour and are UK size 10-12?***".

Analysis of the experimental data revealed that of the user characteristics assessed, spatial ability correlated most strongly with performance (measured as time taken to complete a set number of tasks) on the different dialogue styles. The most striking difference occurred when the users were using the command line interface. Interestingly, the results showed no correlation between personality traits (using the Myres-Briggs Type Indicator) and performance on any of the five different interface styles. These results suggest that at least two types of interface are needed if this particular application is to be suitable for a wide range of users. There is the need for an interface with a dialogue style which minimises navigation and constrains the dialogue for users with low spatial ability, and there appears to be a need for a dialogue style which allows for a more open and flexible dialogue for users with high spatial ability.

Other factors which could have influenced the performance differences in Jennings and Benyon's study must also be considered. The role of short term memory, for example, has not been fully examined in this study. Different dialogue styles may place different demands on short term memory which could result in "cognitive overload" and this may have affected users' performance.

When extending the work of Jennings and Benyon to a spoken user interface involving a similar catalogue task, the first thing that becomes apparent is the difference in the way information is presented to the user in the two systems. In the Jennings and Benyon study, information is screen based, being displayed as written text, icons, graphics and cursor displays. This is in contrast to the telephone dialogue system where information is presented serially in the form of spoken messages and where the user is more limited in the control devices at their disposal i.e. a keypad with twelve keys and / or limited voice recognition, as opposed to the users of screen-based interfaces who have access to a full keyboard, including function keys and pointing devices such as the "mouse". As regards the effects of the interface on the user's memory, in the screen-based interface the user will have a short term memory aid in terms of the information being displayed retained on the screen whereas in using the speech interface will have no such help. Therefore, the designer of spoken dialogue systems should be aware of how individual differences place certain limitations on the individual's processing capabilities and try to accommodate for these differences in the design process.

The relationship found in Jennings and Benyon's study between spatial abilities and navigation within hierarchical structures has also been noted in other studies. Vincente, Hayes and Williges (1987), tested a group of inexperienced users on 21 predictors of individual differences which included demographic measures, verbal abilities and spatial abilities. In this study the subjects had to search for target lines (i.e. look for specific lines in a file) in a hierarchical filing system, using a touch key pad. This allowed them to use file manipulation commands (such as "file select"), large movement commands which allowed string searches (e.g. "search-and") and small movement commands which allowed the subject to scan the current file they were using (this involved using commands like 'page up'). The results showed that spatial ability accounted for more variance in performance than any of the other factors tested. Verbal ability accounted for the second. This study showed a difference in the command selection strategies between subjects of low and high spatial ability. The performance of subjects with low spatial ability seemed to indicate that they were lost in the hierarchical system at times and this was highlighted, in part, by the use of inappropriate search commands. Attempts were made to accommodate these differences by augmenting the interface with provision of a graphical representation of the hierarchical structure of the system and a visual indication of the users' position in the current file they were accessing. These changes reduced the total number of commands issued by users of low spatial ability and also reduced the number of inefficient commands they selected.

Although these strategies worked, there still remained unresolved issues which have direct relevance to telephone dialogue systems. Strategies are required which provide on-line assistance to the user. In telephone based systems it is necessary to provide the user with information about where they are in the dialogue structure, provide help with a specific command query, and provide the opportunity to go back to the top level of the dialogue structure and start the interaction again.

Strategies are also required for adapting the human-computer dialogue; to identify the salient characteristics which affect individuals' performance with a spoken telephone dialogue system; to identify the different types of users and the processes whereby they could choose a dialogue that would be suitable for their requirements; and to ensure that the instructions given on how to use the system should, optimally, be available at any time (via a help command) in the event of the user being in need of a general reminder of how the system operates.

Green, Gomez and Devlin (1986) undertook a similar type of information search study to those already mentioned. In this study, which used novice users as subjects, the researchers were interested in the cognitive factors affecting a subject's strategy for querying a database system. They discovered that reasoning ability was of critical importance when it came to using certain types of interfaces for querying a database. Reasoning was apparently linked to the process of identifying specific targets.

Borgman (1986), undertook a study of undergraduates who were learning to use an on-line library catalogue and produced results which indicate, as other studies of information search tasks have shown, that cognitive characteristics do have an affect on an individual's performance on a computer-based task. In this work, verbal reasoning and spatial ability were identified as the significant factors affecting performance. The study showed that the computer interface did not cater for the needs of certain groups of users: those students who were the most frequent users of the library also had the greatest amount of difficulty when it came to using the catalogue. Therefore the user interface appeared to hinder access for those people for whom it was designed to cater. The study revealed that it was social science and humanities students as opposed to the science and engineering students who had the greatest difficulties, even after being given a training programme. The study also appears to highlight another area which has been investigated, namely the difference in learning styles between science students and social science students.

There have also been studies carried out to examine the role of individual differences in text editing systems. Egan and Gomez (1985) in a study using forty novices, found that the ability to remember the spatial arrangement of objects was a consistently good predictor of user performance as measured by the time taken by the user to complete a set number of tasks. Individuals who scored badly on a spatial ability test made more errors and took longer to perform basic editing operations than individuals who had obtained high scores on the same test. This difference remained after two days of experimenting with subjects using different editing systems and different types of terminals. Other cognitive factors such as verbal ability, reasoning ability and factors such as education or age, were not found to have as significant an effect. The overriding influence of spatial ability was traced to two specific processes in the text editing performance: finding the location of the characters to be edited (in the experiment, the users were given a text containing the changes which had to be made) and the ability to generate a correct sequence to carry out the specified edit. The màjor drawback of this study was that, after identifying the salient characteristics which affect performance, it failed to put forward any solutions for accommodating these individual differences, in the same way as the Vincente and the Jennings and Benyon studies had done for information search systems.

In another study, requiring advanced text editing skills, Gomez, Egan and Bovers (1986) provided what they regarded as further evidence of the importance of cognitive differences in the performance of tasks of this kind. In this study of novice users over a two day period, involving the use of advanced text editing techniques, deductive reasoning ability was found to be a predictor of performance. In this study performance was, as in other studies already mentioned, measured as the time taken to complete a set number of tasks. Card, Moran and Newell (1984) found that there were considerable differences in the editing speed of people in 'technical' occupations as opposed to those in 'non-technical' occupations, fast versus slow typists and frequent versus casual users. In a similar study, involving the evaluation of text editors, Roberts and Moran (1983) found that the average non-technical user was some 15% slower in the performance of error free tasks than the average technical user (once again performance was measured as the time taken to complete a set number of tasks). The study also showed that non-technical users end up spending up to three times longer in error states than the technical user. A similar finding was reported in the study by Vincente et al. However, in the Roberts and Moran experiment, no attempt was made to augment the interface by providing strategies for the non-technical user to accommodate their inefficient use of commands.

2.5.3. Personality Characteristics

From research carried out with human-computer interfaces involving information searches and text editing, there is a body of evidence emerging to indicate that cognitive differences have an important effect on successful performance with such computer tasks.

It has been suggested (Schneiderman, 1980) that as designers of systems explore computer applications in an increasingly widening variety of environments, attention to personality types will be an important consideration. Morgan and Macleod (1990) investigated the role of personality factors in user preferences for computer interface designs in a study which focused on the possible personality differences found between users who preferred to interact with a graphical user interface and users who preferred a command line system. This is an important area for investigation as users are often forced to adopt a particular type of computer interaction method which may not be the one they would choose voluntarily. The experiment consisted of two groups each made up of sixteen subjects (the subject population was university staff and students) allocated to either the command line interface group or the graphical user interface group, depending on which type of interface they used in their normal work.

All subjects were asked to complete the British Standardisation of the 16PF test (Cattell, Eber and Tatsuoka, 1970) and the results showed that there were no significant personality differences between the two groups. This could be attributed, in part, to the relatively small number of subjects (32 subjects) involved in the experiment. Morgan however, highlighted some 'suggestive' trends which were identified by certain factors in the personality inventory. These included the fact that command line users were less sensitive to negative feedback and were more independent, whereas graphical users were perceived as liking security, were more sensitive to negative feedback and were less independently minded. Koubek, LeBould and Salvendy (1985) reviewed three studies which attempted to predict performance from personality scales and found that the correlation between performance and any personality scale was not very strong. In the data produced from their own work, Koubek et al found no significant correlation between performance and personality.

There are however, important aspects which can be considered here in relation to spoken telephone dialogues. It has been noted (Turkle, 1984) that users have a tendency to project personalities on to computer systems. If the user of a spoken telephone dialogue is a sensitive person they tend to perceive the computer (dialogue) in a negative way and this will affect their performance and enjoyment of the system (e.g. they may perceive the fixed tone of the computer as being unfriendly). This suggests that the individual may enter into the human-computer dialogue with the same expectations as they have from human-human conversation. Cox and Cooper (1981) tried to relate judged preference and ascribed personality by asking subjects to rank preference for several computer voices. Each subject characterised the voices by using 28 bi-polar personality scales such as *serene-irritable*. The data were factor analysed and two main factors emerged as significant from the personality ratings: *assertiveness* and *agreeableness*. In general it was found that female voices were just preferred in overall terms to male voices. Those female voices which were perceived as being assertive were more positively rated than the male voices who were perceived as having this trait. These results only provide partial evidence of why particular types of voices are liked and factors such as the nature of the task to be carried out; the content of the information provided in the spoken prompts; and the mode of data entry will also have a profound effect on users' attitudes and expectations. It is worth noting that in this study no attempt was made to examine the possibility of a relationship between the personality traits of the users (and any other individual differences) and their ranked preference of the voices.

2.6. Conclusions

A review of the literature available on the effects of individual differences in human-computer interaction confirms the need to take account of individual differences in the design process of computer systems. Current evidence suggests that cognitive factors such as spatial ability and verbal ability, and other characteristics such as age (Egan and Gomez, 1985), play an important role in determining an individual's performance with computer tasks. When those individual differences that have the greatest impact on the person's use of the system have been identified, this knowledge can then be incorporated in the system design in order to make the interface more usable for individuals with different sets of requirements. This in turn leads to the need to identify the best way to accommodate individual differences in user performance. For the purposes of this thesis, the challenge for spoken dialogue engineering is to develop an experimental framework which will identify the salient individual characteristics affecting user performance with such systems and turn attention towards the accommodation of such characteristics by considering options such as different dialogue interaction styles, in order to make telephone dialogue systems more usable and more easily accessible to a wide range of individuals.

Chapter 3: Methodological Background

3.1. Introduction

The aim of this chapter is to describe the technique developed for the experimental work reported in this thesis and to discuss the Egan and Gomez experimental methodology, which provides the theoretical framework for this work. The background to the Likert questionnaire device used as a measure of usability in the experiments is also provided.

3.2. Wizard of Oz Simulation Technique

The "Wizard of Oz" (WOZ) experimental scheme is a technique for the simulation of human-computer interfaces which is particularly valuable for investigating interfaces currently at the planning stage and therefore not fully operational. The basic idea behind the Wizard of Oz technique is that a "hidden" human, normally referred to as "a wizard", takes on some aspects of the role of the computer in a simulated human-computer interaction.

The origin of the term is unknown, but it is commonly associated with F. Baum's (1900) book - "The Wizard of Oz". In this story, the great Wizard of Oz is a man hiding behind a curtain. This simulation technique is also sometimes referred to as the PNAMBIC (Pay No Attention to the Man BehInd the Curtain) technique and the source of this term has been associated with J.Bernstein (Newell, 1987).

3.2.1. Requirements for Wizard of Oz Simulations

Fraser and Gilbert (1991) counsel that WOZ simulations should not be regarded as a "panacea" and should only be used if certain pre-conditions are satisfied. In the first instance, it must be possible to simulate a future computer system given human limitations. For example, if a future computer system requires the user to navigate through a complex database of information and the navigation tools have not yet been designed there would be little point in simulating such a system. This does not imply however that WOZ simulations are only appropriate for simulating interfaces of applications that are already in existence.

A second important requirement for WOZ simulations highlighted by Fraser and Gilbert is that the designers of the system should have a clear idea of how they expect the future system to behave before they implement any experiments using a WOZ simulation. This is necessary because it is important that the wizard simulates accurately the behaviour of the future system. It is worth pointing out here that although descriptions of speech systems do include error rates, very few indicate the specific type of errors made in sufficient detail to enable a reliable simulation to take place.

Finally, it is very important that the WOZ simulation should convey the illusion that the user is indeed interacting with a "real" computer system. This is easier for some systems than others. For example, if the system only communicates to the user via text displayed on a screen, the wizard may only have to buffer the text to ensure that it appears on the screen one line at a time to maintain the illusion. Simulating a speech output system can however be more problematic. Fraser and Gilbert argue that it is necessary to disguise the wizard's speech to make it sound mechanical. However recent work exploring this issue (Foster et al 1993) has shown that users appear to prefer speech output which consists of the playing of pre-recorded messages rather than the output produced by a speech synthesiser or other voice processor.

3.2.2. Communication Modalities Used in Wizard of Oz Simulations

Fraser and Gilbert (1991) provide a partial taxonomy of Wizard of Oz simulations. An initial division of WOZ simulations can be made into those which use natural language modalities such as speech and typing and those which use modalities such as symbol manipulation or keypad input. Natural language modalities can be sub-divided into those where only one of the participants in the interaction uses natural language and those in which both participants interact using natural language. An example where only the subject uses natural language in a simulated interaction has been reported by Hauptman (1988). In this experiment the subject was asked to manipulate graphical images on a computer screen by issuing spoken commands to the computer. In turn the wizard responded to these commands by typing instructions to the computer which resulted in the images being re-displayed on screen where the subjects had asked them to be placed. Labrador and Dinesh (1984) reported an experiment in which the wizard used natural language while the subject used a non-linguistic modality. In this experiment the subject's task was to access a text-messaging service using a keypad telephone. The system responded to the tone prompts made by the user with synthesised speech output.

In some WOZ simulation experiments both the subjects and the wizard are required to use typed text. Experiments of this type are more straightforward to set-up than experiments which require the use of speech. One reason for this is lack of the class of errors generated during automatic speech recognition which have no correlate in text-based systems where there is a direct correspondence between the key the subject presses and the character which appears on the wizard's screen. One way of minimising this problem in speech input interfaces is to ask the subjects to use a pre-defined vocabulary which will enable the wizard to accept or reject spoken input from the users in a way that is likely to model the performance of the future speech system.

WOZ simulations which require the wizard to type and the subjects to speak are one of the most popular uses of this technique. One example of this type of approach to WOZ simulation is provided by the study reported by Hauptman and Rudnicky (1988) in which subjects were asked to access an electronic mail system using a natural language interface by means of typing or speech to a wizard or using speech to converse with a human operator. The results of this experiment showed that there were significant differences between the way that subjects spoke to a computer and to a human. These results are consistent with the findings produced by Chapanis (1981) in his analysis of human-human interaction. Chapanis found that those modes of communication which require a voice channel are much "wordier" than those that do not. In addition, the number of task-unrelated phrases was again significantly higher for speech as opposed to typed input. Individuals were also more prepared to interrupt a communication system if it has a voice channel. Another example of this type of WOZ simulation is Simpson et al's (1985) database retrieval simulation.

There is a class of WOZ simulation in which both the subject and the wizard interact using speech. Guyomard and Siroux (1988) have conducted a series of experiments of this type based on a simulation of a dial-up Yellow Pages information system. A preliminary data-gathering exercise was necessary since there were no human-human interactions on which to base the WOZ simulation. This resulted in the authors running experiments which used either a constrained, directed dialogue or an unrestricted dialogue. The aim was to obtain data which would provide useful information for designing a spoken dialogue which was acceptable from the user's point of view and was compatible with the then current state-of-the-art for automatic speech recognition technology. The results from this first phase showed that occasional users had problems with the directed dialogue, i.e. this class of users seemed to have problems with producing appropriate answers to yes/no questions. It was also found that in the "unrestricted" dialogues condition the majority of utterances included hesitations and self-corrections.

Richards and Underwood (1984a) conducted a spoken WOZ experiment which simulated a telephone-based railway timetable information service. In this experiment the subjects were told that they would be speaking to a computer and the wizard's (system) voice was distorted to sound synthetic. The subjects also spoke to a "human expert" (who was also played by the wizard). The subjects were informed beforehand that they would either be talking to a person (in one case) or talking to a computer. The results showed that the style and content of the subjects' utterances varied according to the type of system with which they were interacting. For example, when subjects were interacting with the "computer" they spoke more slowly and tended to use a more restricted vocabulary and any queries they had were asked in a direct manner using task-related phrases only.

3.2.3. A New Wizard of Oz Set-up

As can be seen from this discussion of the research literature, most of the WOZ studies of human-machine speech interaction have tended to focus on the problem of characterising dialogue features in specific application domains. The new WOZ scheme (Foster et al 1993) which has been used in the experimental work described in this thesis has been developed specifically to investigate the design, implementation and evaluation of spoken dialogues for automated telephone services. The experimental work reported in this thesis represents the first explicit use of this new Wizard of Oz set-up.

One of the major new features of this new WOZ set-up is that it is based on a realistic simulation of an available speech recognition technology allowing experimentation with recognition levels extrapolated beyond those currently available. In this way the method is able to investigate, among other key issues, the shape of the usability function for automated telephone services for different levels of speech recognition performance. Modifying the speech interface can also involve changing the dialogue used in the automated telephone service. This allows important user interface and human factors issues to be addressed in the experiments such as the impact of voice quality on users' perception of the usability of the service; the degree to which conventional "beep" prompts (used in addition to spoken prompts) influence the progress of the dialogue; and the impact of the dialogue structure and the specific wording of prompts on users' attitude and performance to the service.

The WOZ experimental configuration which includes the subject, the wizard and the simulation software is shown in Figure 3.1.

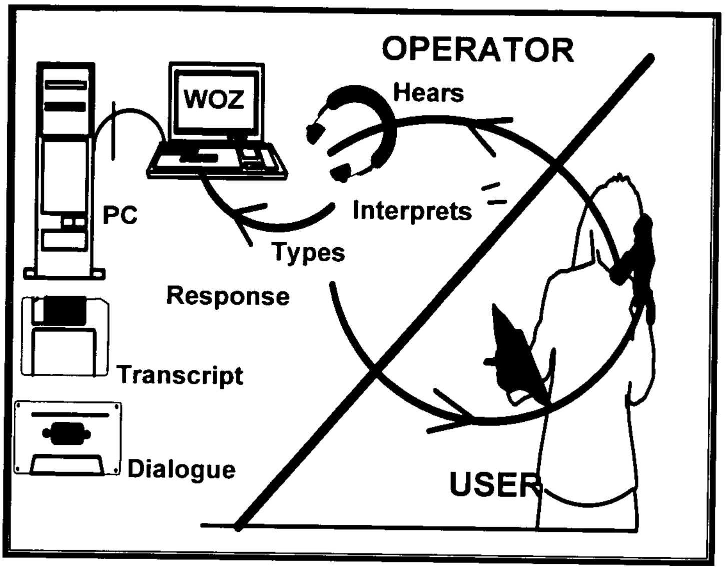


Figure 3.1 WOZ SCHEME

A Software Control Module program handles all aspects of the simulation including the initiation of telephone contact at the subject's home or in the experimental room at the laboratory; the delivery of the pre-recorded dialogue prompts to the subject; the registering of keystrokes from the experimental operator (wizard) made in response to spoken input from the subject; the on-line generation of recognition errors when these are required; and the recording of all statistical data on keystrokes and timings.

The advantages of this particular WOZ configuration include full control over the experimental conditions; constraints imposed on the wizard who only keys in the subject's spoken responses; and the fact that all other experimental factors, such as the introduction of speech recognition errors, are entirely under software control.

A typical experiment using this type of WOZ simulation involves three phases. In phase one (the preamble), the operator (usually the wizard) is connected with the subject to allow normal two-way telephone conversation to establish that the subject is fully prepared to start the experiment. Phase two is the simulated automated telephone service and during this part of the experiment, the wizard's microphone is disconnected leaving only the headphones active. This means that the wizard can hear and key in the subject's responses to the dialogue prompts but any possibility of the wizard leading the subject's responses is removed. In phase three (the signing off), the wizard's microphone is reconnected and the wizard answers any questions the subject may have about the experiment. This could involve the wizard allaying any fears a subject may have about receiving any item they ordered as part of an experiment involving the use of an automated catalogue service.

3.3. Evaluation of Attitudes and Usability

Increasingly in the development of new technology, the attitudes and perceptions of potential users are regarded as important considerations which have to be taken into account in the design and development of systems and services. Users can provide useful information, indicating where the strengths and weaknesses of a system lie. The main problem is that these attitudes and perceptions are difficult to measure. Poulson (1987) argues for a general purpose measuring tool (a questionnaire) which can be used to assess the perceived usability of different systems in various settings.

The aim of this section is to describe a usability metric developed specifically for the evaluation of telephone-based interfaces which has been used as a measurement tool in the experimental work discussed in this thesis.

3.3.1. Usability Metric

Questionnaires and attitude scales are instruments for gathering structured information from people. The aim of the usability research described here is to develop an attitude questionnaire which is a reliable and valid measuring instrument for testing a particular opinion or pattern of behaviour. In order to assess user attitudes towards spoken dialogue systems for the experimental work reported in this thesis, a Likert-type attitude scale was constructed (Likert, 1932). The advantages of using the Likert technique have been identified by Coolican (1994) as:

- Subjects prefer the Likert scaling technique because it is "more natural" to complete and because it maintains the subject's direct involvement

- The Likert technique has shown to have a high degree of validity and reliability

- The Likert scale has been shown to be effective at measuring changes over time

In designing a Likert-type attitude scale there are several important issues which must be addressed. In the first instance, it is necessary to produce a number of statements that explore the range of facets which influence the attitude under test - which in this case is attitudes towards spoken dialogue systems for the telephone. In the questionnaire it is essential to achieve a balance between positive and negative attitude statements in order to overcome the danger of the overall score obtained from the attitude scale reflecting the users' biased tendency to agree rather than disagree with the attitude statements (an effect known as 'response acquiescence set') instead of providing valid information on their attitudes. This is due to the fact that the agree-disagree scale consistently goes from left to right and there is a tendency for people to fill in only one 'position' on the page. For each statement each subject is asked to indicate the extent to which they agree or disagree with it and for the purposes of the work reported as part of this thesis, a seven-point scale with a mid-neutral point was used. Table 3.1 gives an example of the format used for this attitude measurement tool.

The service was easy to use

□	□	□	□	□	□	□
strongly agree	agree	slightly agree	neutral	slightly disagree	disagree	strongly disagree

Table 3.1 Likert format used in Attitude Questionnaire

As stated, the values on the scale range from 1 to 7. If the subject strongly agrees with a positive statement they would score 7 for that particular item but they would only score 1 for strongly agreeing with a negative statement. After a subject completes the Likert questionnaire the scores for each item are summed (allowing for statement polarity reversal) and a mean is taken to provide an overall score for a subject's attitude towards the usability of the automated telephone service.

After completing the Likert statements, subjects answer a series of open-ended questions (e.g. What did you like about the service?) and a general comments section which allowed them to express opinions and perceptions not covered by the attitude scale. Including these provides a source of qualitative data which can augment the quantitative data provided by the attitude scale.

Having designed an attitude measurement device the next step is to assess its reliability and validity. Rust and Golombok (1989) define reliability as the extent to which a test accurately produces the same results on different occasions and validity is defined as the extent to which a test actually measures what it claims to measure. The validity of the usability questionnaire was assessed by a process known as content validity. This method of assessment requires researchers with expertise in the required area (in this case the usability of automated telephone services) to evaluate the content of the measurement tool in order to assess the extent to which it is representative of the area it is supposed to cover. An emphasis is placed on ensuring that all the important elements in the topic area are adequately covered without any undue weighting being given to some aspects of usability compared with others (Foster et al, 1993). The aspects of reliability and validity have been addressed throughout the series of experiments discussed in this thesis.

3.4. Background to the Egan and Gomez Approach

Interacting with a computer is a complex task. Computers vary in their functionality, purpose and structure People differ in terms of personality, cognitive skills and gender. Computer systems are designed to provide an effective and satisfying interaction for a wide variety of users, yet, as Benyon (1993) points out, there have been few attempts to consider what effects individual differences have on human-computer interaction. The three step approach offered by Egan and Gomez represents the first systematic approach to understand the effects of individual differences in human-computer interaction. As result of this, other researchers who have been interested in this area of research (e.g. Benyon, 1993, Vicnete, Hayes & Wiilges, Jennings, 1991) have employed the Egan and Gomez philosophy in their own experimental work.

Traditionally, the study of individual differences has been carried out in order to predict what kind of people will perform well for a given task. The approach of Egan and Gomez differs from that traditional approach in its goals and methods. Their goal (based on the three steps of **assaying, isolating** and **accommodating**) is to find a way which allows a range of people to acquire a complex skill that may be part of a task. The emphasis is on *accommodating* people through systems engineering design and training rather than finding more efficient ways of selecting individuals for a given job. They also highlight the need for identifying groups of individuals who are having more or less difficulty in learning a skill, as a first step to focusing attempts at system design and training.

The methods adopted by Egan and Gomez also differ from the conventional approach where a predictor variable (e.g. verbal ability) can be "opaque" and yet can still be considered useful because it accounts for a certain percentage of unique variance in a criterion measure. No emphasis is placed on trying to understand why a particular variable is related to task success. As a result there is a tendency to use composite measures such as quick-scoring intelligence tests as predictors of success in conventional tasks. The Egan and Gomez approach, in contrast, looks at variables which assess specific cognitive capacities. They believe that difficulty in learning a complex skill - the example they have based their work on has been computer text editing - can be predicted from a small set of individual characteristics which are clearly understood. This goal may not always be achievable but Egan and Gomez regard it as the optimal outcome of the "assay" of individual differences. Emphasis is placed on looking at simpler variables because they can be linked more easily than composite measures to manipulations which could influence their importance. For example, it is noted that people of higher intelligence can learn a text editing task faster than people of lower intelligence. However this does not offer any suggestions on how to change the conditions to make learning easier. If, on the other hand, the experimenter has gathered evidence which shows that learning a text editing task correlates with a specific capability (e.g. verbal ability) this could provide useful information on how to improve the performance of say the low verbal ability subjects by improved system design.

The experimental approach of Egan and Gomez incorporates two well-known experimental approaches to the understanding of cognitive differences. The first of these is the cognitive correlates method (Hunt, Frost and Lunnebourg, 1973) which tries to explain individual differences in a complex ability such as verbal ability by relating the differences found in this ability to differences in simpler information-processing measures such as the time taken by subjects to make a physical name match of alphabetic characters. Egan and Gomez propose to use the same techniques in their method but with attention being focused on how to understand a particular skill rather than a general ability.

The second method adopted by Egan and Gomez is component analysis (Sternberg, 1977). This refers to a form of analysis which is based on the determination of components or distinctive elements. In the experimental work reported by Egan and Gomez the idea was to break-down the skill of text editing into a series of component operations and identify the individual differences that correlate with each of the task components in order to improve their understanding of how individual differences can affect performance in a text-editing task.

Egan and Gomez (1985) chose computer text editing as a case study as this task provided the potential to produce large individual differences in learning difficulty, and to offer interesting possibilities for accommodating individual differences in this skill.

3.4.1. Step 1: Assaying Individual Differences

The first step of the Egan and Gomez approach is to discover the important sources of individual differences in the performance of a task. They conducted two preliminary studies. In the first study a teacher and four students from a course on computer text editing were interviewed. Two of the students had performed extremely well on the course and two of the students had performed poorly on the course. The students and teacher were asked how successful or unsuccessful students might potentially be distinguished before the start of the course. They were also asked what they found difficult about the course and what kind of preparation students should have for the course.

Potential correlates of the difficulty of learning text editing emerged from these interviews. Successful learners were described as being confident and motivated in their approach to computers as well as being willing to try things on their own. Unsuccessful learners were not seen as having these qualities. People's experience of and attitudes towards machines and mechanical devices was also found to be a predictor of success. Learner's age was also found to be a possible correlate with success, with older subjects having more difficulty with the course than younger ones. Finally, a good memory may be required to deal with the different editing commands.

From these interviews Egan and Gomez constructed a questionnaire to assess potentially important learner characteristics. The questionnaire asked subjects' their age and typing speed amongst other information and assessed subjects' attitudes towards computers using Likert-type statements, e.g. "I am afraid of computers". Another section assessed subjects' experience and desire to use computer-like devices (e.g. automated bank tellers).

The second preliminary study consisted of six subjects who completed the questionnaire and who gave step-by-step explanation of their actions as they worked through a tutorial on text-editing. As a result of this, there was a modification of the questionnaire and the training materials which were subsequently used in their first experiment.

The first experiment assessed learner characteristics which were those measured by the revised questionnaire and three standard psychometric tests. These tests were chosen because they were reliable measures of the distinct characteristics which might be important in learning text-editing skills. The Controlled Associations Test (Ekstrom, French and Harmen, 1976) measures the fluency with which subjects can generate semantic associations. Associational fluency was regarded as having potential importance in verbal processes such as understanding and remembering text-editing commands. The Building Memory Test (Ekstrom, French and Harman, 1976) measures subject's ability to remember the spatial arrangement of objects. Spatial memory was regarded as having the potential of playing an important role in processes which involve visualising changes made to a piece of text, finding the location of changes to be made in a text. The Nelson-Denny Reading test (1973) was chosen to measure subjects' reading skill. The overall score was made up of vocabulary and reading comprehension sub-scores. Reading skills were seen as being potentially important since the text-editing tutorial involved comprehending written instructions.

Thirty-three adult women participated in this first study. All were classed as being naïve computer users. The subjects were given a text-editor tutorial to complete. This consisted of a tutorial manual and an interactive program that presented exercises to be completed after each section of the manual. The manual listed six text-editing operations:

- adding lines of text
- removing lines of text
- inserting characters within a line
- removing characters within a line
- replacing lines of text
- exchanging one character string with another within a given line

The subjects were asked to use three text-editing commands: append, delete and substitute. There were six sections to the manual, each describing how to use an operation. The subjects read each section and then performed an exercise they had just learned. A complete record of the subject's interactions with the computer was obtained including every keystroke made by the subject.

This complete record of subjects' interactions with the computer allowed the authors to select three primary performance measures (reading time, execution time per successful change, first-try errors) which would characterise the difficulty of learning text-editing skills.

The results from this experiment showed that there were wide variations in the performance measures which seems to confirm the belief that there are sizeable individual differences in a task like this. Reading time for tutorial materials was found to be uncorrelated with first-try errors although it did significantly correlate with execution time per successful change ($p < 0.01$) which may suggest that part of the execution time may have been spent on re-reading the manual. Execution time also significantly correlated with first-try errors ($p < 0.01$). This suggests that the time spent by a subject on making an editing change depended on whether or not their first attempts at interacting with the system were correct.

On the basis of their experimental findings, Egan and Gomez selected only three individual characteristics (Building Memory Test score, Nelson-Denny test score and age) for further analysis. These all correlated significantly with reading time, execution time per successful change and first try errors. These were in fact, with one exception, the only predictor variables which correlated reliably with any of the three criterion measures.

For each of the performance measures a multiple regression equation was constructed giving the best-fitting linear prediction based on reading skill, spatial memory and age. The results from this indicated that each equation accounted for a sizeable proportion of the corresponding performance measure. Egan and Gomez regarded the pattern of regression coefficients produced by the multiple regression analysis as having a straightforward interpretation. For example, reading skill was seen as being able to predict aspects of learning difficulty that directly involve reading, i.e. reading time and execution time per successful change. Spatial memory and age are strongly related to the part of learning that involves interacting with the computer. These two variables are the only ones which contributed reliably to the prediction of first-try errors and like reading skills make significant contributions to the variance in execution times per successful change.

Overall, two important results emerged from this study. Firstly, a large variation in subject's difficulty in learning text-editing was identified which appears to be systematic. Secondly, a large amount of this variability can be accounted for by a few distinct individual characteristics. In addition, not all the characteristics of the individuals were found to have equal importance in predicting performance. Most of the characteristics assessed such as verbal ability, previous experience and attitudes towards computers did not significantly correlate with any of the criteria chosen. Instead, time spent reading was predictable primarily from subject's scores on the Nelson-Denny Reading Test and time spent interacting with the computer and the number of errors made during this interaction were predictable from spatial memory and age.

Egan and Gomez conducted a second study which was aimed at replicating and extending the results obtained from the first experiment. This experiment was also designed to look at the learning that occurred as subjects worked through the tutorial. The first experiment did not allow for a precise assessment of learning and the results obtained from that experiment could have been due to the novelty of the users' first interaction with a computer. In addition, it is possible that the predictors of an individual's ability in a text-editing task could change as learning progressed. For example, typing speed was not regarded initially as being important but may have become important as the user became more experienced. To develop a better assessment of learning and the correlates of performance after some practice, subjects were given two tutorials to complete, separated by a week, in the second experiment. The study was also designed to confirm that age and the spatial ability measure predicted learning difficulty. For example, age could be important as it could be correlated with educational achievement, recency of education or perhaps with attitudes and experience with computers. Spatial memory may be correlated with difficulty in learning text-editing because memory is in general important, or because the Building Memory Test involves strategies akin to reasoning. In the first experiment, no data was provided on the covariates of age and spatial memory.

To try to control for this, Egan and Gomez assessed several potential predictors of text-editing skills in addition to age, spatial memory, reading ability and age:

- number of years in education
- attitude to computers
- associative memory capability
- logical reasoning capability

The addition of these other predictors allowed Egan and Gomez to investigate whether age and spatial memory made unique contributions to the prediction of learning difficulty in a text-editing task.

The results from this second experiment, which used a similar procedure to that of the first experiment, indicated that subjects did learn from their first interaction with a computer and remembered enough to improve their performance a week later. The experiment found that the three performance measures (reading time, execution time per successful change and first-try errors) showed a 30% - 40% improvement for the second trial compared to the first. However despite this overall improvement, all three measures continued to show a variance in performance for the second exposure (e.g. some subjects still made errors 50% of the time they first tried a problem).

Analysis of the predictors of text-editing difficulty showed that spatial memory and age were again the best predictors. Each of these measures significantly correlated with first-try errors and execution per successful change on both days. A forward stepwise multiple regression was used to investigate which variables made reliable and unique contributions to predicting user performance on the task. All the predictor variables that significantly correlated with either first-try errors or execution time per successful change on either day were candidates for inclusion in the regression models. Forward stepwise regression involves entering the variable which accounts for the most variance in performance into the equation first. Other variables are only entered if they make a significant contribution to the variance in performance. The results from this analysis showed that age and spatial memory were the only variables to enter the equation for each performance measure. A backward stepwise regression analysis where all candidate variables are entered and are then removed one by one if they do not make a significant contribution to the variability in performance produced the same result as the forward stepwise analysis.

This experiment also allowed classification of first-try errors and examination of the correlates of various error types. Error analysis was viable in this case because the nature of the task required each subject to make a large number of editing changes with each exercise being repeated at a one week interval. By breaking down error types it was possible to show correlations between subjects' individual characteristics and the total errors made in each category. This suggests more precisely how age and spatial memory affect performance. For example, spatial memory test scores are seen as being a good predictor of the frequency of pattern errors such as when the subject does not produce the correct sequence of characters and symbols for a command. Age was seen as being a good predictor of omission errors when elements of a command are missing.

To sum up the findings of these first two experiments, the difficulty of learning to use a text-editing system have been shown to be related to subject's individual differences in terms of age and spatial memory. The kind of information this provides may be of practical use for teachers in that they can anticipate the amount of difficulty that different groups of students would have when they are learning to use a computer text-editing system. In order to accommodate for these differences in learning and spatial memory, Egan and Gomez felt they would have had to go further than the correlation results they had obtained from the first two experiments.

3.4.2. Step 2: Isolating Individual Differences

Egan and Gomez developed their approach by breaking down the text-editing task into smaller components. The purpose of this approach was to identify those components of the task which accounted for the most variance in text-editing difficulty and to determine which components are the most sensitive to subject's age and spatial memory. It was felt that if this kind of information was obtained it would be easier to pinpoint those aspects of text-editing that could be made simpler for the subjects to use through software design, training or performance supports.

Egan and Gomez proposed that a typical text-editing operation could be decomposed into 3 general components:

(a) **Finding**: this is the process of locating the place where a change has to be made in the computer's version of the text.

(b) **Counting**: this is the process whereby subjects obtain the line number for an editing change. In these experiments the subjects had actually to do this, they did not have things like context search or line number indicators.

(c) **Generating**: this is the process of producing the correct sequence of symbols and patterns to accomplish the desired editing change.

These component processes were assessed using simulated tasks of specific parts of the computer text-editing task. This was carried out using a pencil and paper format and included a large number of items which required one, two or all three component operations. The idea was that subjects had to complete as many simulated items as they could within a given time limit. This method was chosen because it was believed that very few subjects would complete all of the necessary operations within the given time.

The results of this experiment indicated that performance on the component tasks was comparable to actual text-editing performance. In particular it was found that the finding and generating components correlated almost as strongly with actual text-editing difficulty as did the complete task. No significant results were found for the counting component. In addition, out of the three single components, generating produced the highest correlation with actual text-editing measures.

Subject's age and spatial memory test scores were used in multiple regression equations to predict each component task process. The results of this analysis indicated that age had a significant effect in predicting every process involving the generating component only. This could be due to the fact that older people experience greater difficulty when generating complex editing commands. Spatial memory had a significant correlation coefficient for every process which involved either finding or generating. Spatial memory has, it would appear, an important role to play in more than one component of text-editing.

These results led Egan and Gomez to develop a working hypothesis which suggests that the effect of age is concentrated in the generating process while spatial memory has an effect on both the finding and generating processes.

3.4.3. Step 3: Accommodating Individual Differences

As spatial memory has been implicated in both the finding and generating components, this ability may also have a significant effect on other complex skills which require similar component processes. This leads to the question of how to reduce the demands on spatial memory with computer-based tasks of this kind. There are already available certain performance aids which have been designed and implemented to support the finding process. For example, the availability of a context search facility may reduce the amount of effort placed on an individual's spatial memory. Another way to reduce this type of cognitive loading would be to develop an "intelligent verbal editor", which eliminates the visual finding process by recognising the spoken verbal corrections made by the user and locating the appropriate piece of text to be changed.

Another factor to be considered is that the generating process defined by Egan and Gomez is a process which is frequently used in human-computer interaction. They point out that in a text-editing task, if the user experiences difficulty in remembering the spatial pattern required for the generating of a command, the spatial pattern could be produced first in order to reduce problems with memory due to interference or delay. Thus a finding - generating - counting sequence may not place as great a burden on spatial memory capacity as the standard finding - counting - generating sequence. Having the generating process immediately following the finding process reduces the effect of delay or interference on the storing and retrieving of spatial patterns. Another factor which has to be accounted for in system design is whether or not performance is affected by user's age. From the results reported by Egan and Gomez it appears that older people experience greater difficulty when it comes to generating complex command sequences in text-editing tasks. This result correlates with other age versus complexity studies which indicate that general measures of a system's complexity, such as time duration, can be used to predict how sensitive the system is to user's age.

In terms of their approach to accommodating individual differences in a text-editing task, Egan and Gomez focused on different training procedures rather than changing the design of the text-editing system. One approach which was offered to help people with poor spatial memory was to provide training in processing strategies. This involved the user completing the counting component before trying to remember the pattern of characters required. Secondly, users could go back to the original copy of the text to obtain the old set of characters that are required to be substituted and then going back, in a separate step, for the replacement string. This should reduce output delays and interference and result in fewer pattern errors being made.

In order to make learning less age sensitive, Egan and Gomez propose avoiding, in the first instance, trying to learn a complex procedure such as the string substitution command. Instead, subjects could be taught how to change one full line of text for another by combining the delete and command functions they have learned already or by learning the relatively simple change line command. This takes into consideration the fact that commands that operate on entire lines are usually easier to generate than character-string substitutions. Egan and Gomez believe that the overall complexity of a text-editing system can be reduced if users learn these more straightforward substitution commands in the first instance.

3.4.4. Summary of the Egan and Gomez Approach

The approach proposed by Egan and Gomez has been adopted by other researchers interested in dealing with the effects of individual differences in human-computer interaction (e.g. Benyon , 1993, Vicente, Hayes & Williges, 1987, Jennings, 1991).

There are however potential problems to this approach. For instance there could be problems in isolating where individual characteristics exert their greatest influence. A specific analysis of a complex task will not necessarily be the uniquely correct way of decomposing the particular task. Various task analyses may differ in their interpretations of component processes and the level of detail in which they describe the processes. One way of dealing with this potential problem, offered by Egan and Gomez, is to look at the usefulness of the task analysis that has been employed. A useful task analysis is one which identifies the component processes that are responsible for the individual differences in a complex task and identifies these as components which can be changed, removed or supported in some way. If the emphasis is on understanding individual differences in order to accommodate them in some way, this could improve the chances of producing a more useful task analysis.

Once individual differences have been assayed and isolated there is still the problem of accommodating the appropriate changes. The accommodation stage tests the hypothesis created by the two steps which precede it and also tests the theory of how an experimental manipulation, i.e. new design or different instructions, will change the nature of the original task. For example, spatial ability is strongly correlated with a particular aspect of performance on a complex task. The question of how to deal with this is a major question. It is possible to support performance with simple diagrams which will display relationships which people with low spatial ability may have difficulty in understanding and imagining.

Alternatively, the designer could dispense with any form of spatial reference. The point is that even after extensive prior analysis, the ways to accommodate changes in a complex task are not always apparent. Egan and Gomez counter this by pointing out that attempts to accommodate individual differences in a complex task after going through the assay and isolation stages have a better chance of succeeding than attempts which have been undertaken without this prior analysis. As evidence for this claim they cite a series of studies (Cronbach & Snow, 1977) which highlight the difficulty in trying to accommodate for individual differences without the use of any prior knowledge of which individual differences are important or why. Without adopting the three stage approach of Egan and Gomez, identifying the salient individual characteristics and understanding why, for instance, one particular form of editing produced better results than another would have to rely to a great extent on guesswork.

The work of Egan and Gomez has shown that salient individual differences in learning a complex task can be assayed, those components of the task which cause some people undertaking the task greater difficulty can be isolated and these task components can be changed to accommodate a larger number of people who use the system. Instead of adopting the traditional approach of individual differences in which the emphasis is on selecting the 'right person' for the 'right job', Egan and Gomez propose developing the 'right conditions' (i.e. user instructions, training, performance aids) which will allow a specific individual to accomplish their desired objective.

The underlying promise of this approach is that it could prove to be most useful in situations where a large number of people from different backgrounds wish to learn a new complex skill and for which there exist possibilities for accommodating individual differences. This approach is judged to be especially relevant for studies in interactive telephone services since the telephone is increasingly being used to access databases (such as flight information services), home banking and home shopping services. This has exposed a great variety of people to new technology for the first time and as a result of this the design of the interface between users and automated telephone services has become a major focus of research and developmental activity.

The aim of the dialogue engineer is to design a dialogue service which provides an interaction that is comfortable and satisfactory for the user as well as being efficient. To achieve this, information on the users' capabilities must be fully assessed in the design phase. The approach offered by Egan and Gomez provides a way for highlighting those individual characteristics which have a significant effect on user performance and perception of the usability of telephone-based interactive dialogue systems.

Chapter 4: Assaying Individual Differences for Interactive Telephone Services

4.1. Introduction

The aim of this thesis, as outlined in Chapter One, is to investigate the effects of individual differences on user performance with spoken dialogue systems for the telephone. This Chapter reports on the first step (assaying) of the three-stage approach outlined in Chapter Three to study the effects of individual differences on user performance when interacting with a spoken telephone dialogue system.

The work reported in this Chapter is based on an experiment specifically designed to investigate the relationship between individual characteristics (such as personality type, cognitive skills and age) and subjects' performance on different input modes for an automated telephone service. The experiment was intended to show that differences in subjects' performance on the three input modes can be attributed to individual characteristics and that subjects' performance when interacting with spoken dialogue systems can be predicted on the basis of these salient individual characteristics. The dialogue designed for use in this experiment allows subjects to interact with a simulated home shopping service. This service involves users in three major tasks, as McInnes (in press) details. The dialogue begins with a message from the automated catalogue welcoming the user. After this, subjects proceed to the first phase of data entry which requires them to enter an 8-digit customer identification number, segmented into two blocks of 4 digits. The number recognised by the service is read back for subjects to confirm or reject and if the number recognised by the service is incorrect, this first data entry phase is repeated. The second data entry phase requires subjects to input an 8-digit item number taken from a shopping catalogue, again segmented into two blocks of 4 digits. After the catalogue item number has been entered, the system reads back a description of the item from the catalogue together with its price and once again a yes/no confirmation response is requested from the user. If the catalogue item number given by a subject does not correspond to any item available in the catalogue, the service plays a message stating this. If the subject responds with a negative response to the confirmation question the catalogue item number entry phase is repeated. The third and final data entry phase is the entry of a 16-digit credit card number segmented into 4 blocks of 4 digits and read from a replica credit card. Once a user has entered their credit card number the service reads this number back and again a yes/no confirmation response is requested from the user. The transaction is completed with the service providing the user with a description and price of the item ordered from the automated catalogue and a final signing off message.

The three input modes used for interacting with the automated telephone system in this experiment were chosen to be representative of those available with current spoken telephone dialogue systems: *isolated word, connected word* and *keypad* input.

In the *isolated word* interface the user is instructed by the service at the start of each block of digits to say the number one digit at a time after the 'beep'. The service provides users with this prompt at the start of each block of digits, while the prompts between digits within a block consist only of 'beeps'. If a user speaks over the 'beep' or says something outwith the permitted vocabulary (the accepted vocabulary for digit entry consists of the words 'one' to 'nine', 'zero', 'nought' and 'oh') the Wizard of Oz operator will hit a 'reject' key and the service will play an error message ('Sorry, I didn't understand that'), after that the user will be prompted by the service to input the digit again. If the user does not input a digit or yes/no confirmation following a prompt by the service, after a predetermined time-out period (5 seconds for the simulated automated catalogue service used in this experiment) the service will respond with a suitable prompt (e.g. 'Sorry, I didn't hear anything. Please say the first 4 digits of your credit card number, clearly saying one digit after each beep'). After the user has entered a block of four digits, the service reads back the recognised digits to the user and looks for confirmation that the digits it recognised are correct; if the user says 'no'; the service will request the user to enter the block of digits again.

The *connected word* interface is similar to the *isolated word* interface except that each block of four digits is spoken connectedly by users after a single prompt from the service ('Please say the first four digits of your personal identification number, speaking clearly after the beep'). If the subject inputs fewer than four digits by the end of the time-out period, or if a non-digit input is given, the service will reject the input and the user will be prompted for the block of four digits again.

In the case of the *keypad* interface, there is a single prompt for the customer identification number, the catalogue item number and the credit card number. After this, the user has to enter the number as a single sequence of 8 or 16 digits using the telephone keypad. The keypad version of the service does not have a 'beep' and the user can interrupt or 'type through' the prompt. If the user does not enter the correct number of digits or presses an invalid key (e.g. user presses square key '#' instead of a digit key) the service will reject the input and prompt the user to enter the number again. In this version of the automated catalogue service, the confirmation prompt asks the user to press '1' if the information read back by the service is correct or '2' if it is not.

All three versions of the automated catalogue service allow users up to three attempts to enter a digit or a digit sequence. If a user's third attempt at digit entry is rejected by the service, the dialogue fails; the service informs the user that there appears to be a problem and the user is returned to the wizard. In addition, the dialogue specifications for the isolated word and connected word input modes required the simulated word recogniser to be set at 100% recognition accuracy to enable a valid comparison between the speech and keypad input modes.

Subjects' performance was measured by call duration (i.e. the length of time taken to complete the task), error rates (i.e. the number of errors made by subjects when interacting with the service) and with the Likert usability questionnaire. These measures are similar to those suggested by Shneiderman (1987) in Chapter One as being suitable for quantifying the efficiency and usability of an interactive system.

4.2. Methodology

4.2.1. Experimental Design

A repeated measures design was used with six balanced groups each consisting of 14 subjects, with each group representing one of the six possible orders of exposure to versions of the service, e.g. connected word then keypad then isolated word. Each subject used the automated catalogue service three times, each time with a different input mode.

4.2.2. Subjects

In total, 84 subjects from a local subject panel took part in this experiment. Subjects had no previous experience of spoken language dialogues for automated telephone services. The characteristics of the subjects are given in Table 4.1.

	Ages 18-38	Ages 39-59	Age 60+	Total
Male	16	10	12	38
Female	16	20	10	46
Total	32	30	22	84

Table 4.1 Age and Gender Breakdown of Subjects used in Automated Catalogue Service Experiment

4.2.3. Materials

In the first part of the experiment all subjects completed two psychometric tests. The NEO-PI-R is a five factor test consisting of 240 personality questions. This test, which is not timed but typically takes between 45 minutes and 1 hour to complete, is a concise measure of five major traits of personality: neuroticism, extroversion, openness, agreeableness and conscientiousness (see Chapter Two). The second test that subjects completed was the AH4 Group Test of General Intelligence involving two 10 minute timed tests, one for verbal ability and one for spatial ability.

In order to use the automated catalogue service each subject was given a customer identification number, a shopping catalogue and a replica credit card. The subjects were given a Likert attitude questionnaire to complete after each use of the service (but not after they had used the service for a fourth time on a free choice option).

4.2.4. Apparatus

The Wizard of Oz experimental workbench described in Chapter Three was used in the experiment. For the purposes of the experiment reported here, an IBM-compatible Personal Computer with a plug-in board (which provides a connection to the telephone line) was set up in an office near the laboratory in order to run the WOZ software. A telephone was also provided in the laboratory for the subjects to use as part of the experiment.

4.2.5. Procedure

Subjects first completed the NEO-PI-R personality inventory. Immediately following this, subjects completed the AH4 Group test of General Intelligence. After a short break, they used the automated catalogue service three times, each time with a different input mode. Following each use of the service, subjects completed a 32-question Likert attitude and usability questionnaire. After the subjects had used the automated catalogue service three times, they were asked to select their most preferred version for a fourth use.

Finally, subjects were interviewed for about 10 minutes regarding the different versions of the service they had used as well as their familiarity with and attitudes towards automated telephone services, computers and technology in general.

4.3. Results

Two objective measures of performance were selected. The first measure was duration, the total amount of time a subject spent completing the task of ordering an item from the automated catalogue. This measure was obtained by calculating the time from the wizard hitting the key labelled "go" to start the human-machine interaction until the prompt handing the user back to the human operator, which ended the human-machine interaction.

The second measure of objective performance used in this experiment was error rate. This measure was defined as the number of errors a subject made when interacting with the spoken language dialogue interface. Errors are defined as user silences - the failure of the subject to speak after being prompted by the system and system rejects - where a user response is rejected if it is a non-vocabulary item e.g. saying "nil" instead of "zero", or if the user speaks over the beep prompt.

In addition to these two objective measures of performance, the Likert attitude and usability questionnaire provided subjective data on a 7-point scale of subjects' attitudes with respect to the three interfaces.

4.3.1. Overall Results

The first data to be analysed were the overall mean call durations for the three input modes. Table 4.2 shows the mean call duration for all subjects across the three input modes.

Input Mode	Mean Call Duration
Keypad	113.92
Connected Word	156.36
Isolated Word	172.49

Table 4.2 Mean Call Duration (in secs) for All Subjects Across Three Input Modes

A Related T-Test indicated that there was a significant difference in time taken to complete the experimental task when keypad and connected word input modes are compared; {t = -20.48, df = 83, p < 0.001}; when keypad input and isolated input modes are compared, {t = -.22.09, df = 83, p < 0.001}; and when connected word and isolated word input modes are compared, {t = 5.78, df = 83, p < 0.001}. These results are not surprising given the differing speed with which an individual can enter digits using these three input modes. The next overall result to be considered was error rates. Table 4.3 shows the mean number of errors for all subjects across the three input modes.

Input Mode	Mean Number of Errors
Keypad	0.72
Connected Word	0.33
Isolated Word	0.62

Table 4.3 Mean Number of Errors for All Subjects Across Three Input Modes

A Related T-Test indicated that there was a significant difference between the number of errors subjects made when using the keypad input version of the automated catalogue service compared with the isolated word input mode, {t = -2.81, df = 83, P < 0.01}; and when the number of errors made by subjects using the connected word and isolated word input modes are compared, {t = -2.13, df = 83, p < 0.05}. No significant differences were found between the number of errors subjects made when using the keypad and connected word input modes. The significance of these results are explained in more detail when the errors made by subjects when using the three input modes are analysed in terms of individual differences.

Subjects attitudes towards the three versions of the automated catalogue service were then analysed. Figure 4.1 shows the overall mean subjects attitude scores for each of the three input modes.

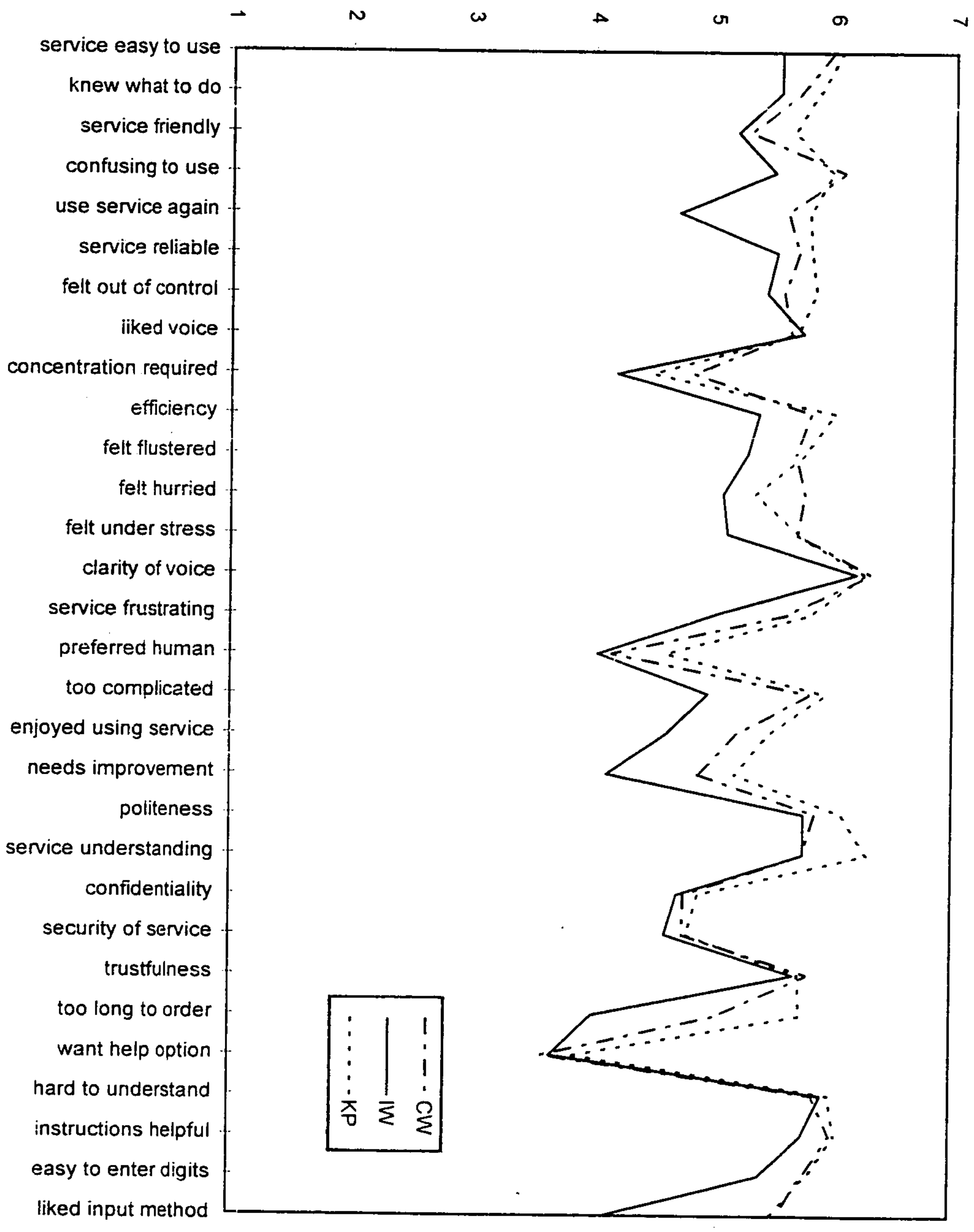


Figure 4.1
Mean Subject Attitude Scores Across the Three Input Modes

The mean attitude scores for the three input modes were analysed and a Related T-Test found a significant difference in subjects attitude to the keypad input and isolated word input version of the automated catalogue service,
{t= 5.35, df = 83, p < 0.001}; and between subjects attitude to the connected word input and isolated word input versions of the service, {t = 5.90, df = 83, p < 0.001}. No significant difference was found between subjects attitude towards the keypad and connected word input versions of the service. The attitude scores for the three input modes were analysed in more detail by considering the possible effects that order of use may have on subjects perceptions of the usability of the three input modes. Table 4.4 gives the mean attitude score in terms of order of use for the keypad, connected word and isolated word input modes.

Input Mode	Order of Use		
	1	2	3
Keypad	5.7	5.51	5.59
Connected Word	5.56	5.37	5.46
Isolated Word	5.53	4.79	4.95

Table 4.4 Mean Attitude Score for Order of Use

A series of One-Way ANOVAs taking order of use as the categorical variable and mean attitude as the dependent variable for each input mode indicated that there was no significant effect for order of use on attitude to the connected word or keypad input versions of the automated catalogue service. There was however a significant effect for order of use on subjects attitude towards the isolated word version of the automated catalogue service (p < 0.01). If one looks at the figures shown in table 4.4, the isolated word input mode received the same mean attitude score for novice users as the connected word input mode and only slightly lower than that given to keypad input. This result indicates that users with no previous experience of automated telephone services adopt very similar attitudes regardless of input mode. However, as soon as subjects ceased to be novice users - on their second and third uses of the automated catalogue service - isolated word input was considered to be significantly less usable than either the connected word or keypad input modes. This result suggests that users who may be familiar with either a connected word or a keypad input service would adopt a negative attitude towards a new service, such as the automated catalogue service, if it was only offering isolated word as the input mode for interacting with the service.

4.3.2. Individual Differences in Cognition - Objective Measures Analysis

The scores for the verbal and spatial sub-scales of the AH4 Group Test of General Intelligence were grouped into two categories - high or low - the former being defined as the upper half of the sample distribution scores for each of the sub-scales and the latter being the lower half of the sample distribution. Table 4.5 shows the mean scores for call duration for connected word, keypad and isolated word input modes for low and high verbal ability subjects.

Input Mode	Low Verbal Ability	High Verbal Ability
Isolated Word	172	173
Connected Word	161	151
Keypad	120	108

Table 4.5 Mean Call Duration (in secs) for Verbal Ability

A One-Way ANOVA (Related) was performed for each input mode taking mean call duration as the dependent variable and level of verbal ability as the categorical variable. Significant differences were observed between the high and low verbal ability groups on the connected word interface, {$F(1,82) = 9.423$, $p < 0.01$}, and the keypad interface {$F(1,82) = 12.114$, $p < 0.001$}. No significant difference was found between high and low verbal ability subjects when using the isolated word interface.

Individual Difference #1

Mean call durations are significantly longer for low verbal ability users than for high verbal ability users when using connected word or keypad input to complete a complex task.

The mean scores for call duration for connected word, keypad and isolated word input modes for low and high spatial ability subjects are shown in Table 4.6.

Input Mode	Low Spatial Ability	High Spatial Ability
Isolated Word	173	172
Connected Word	161	152
Keypad	122	107

Table 4.6 Mean Call Duration (in secs) for Spatial Ability

A One-Way ANOVA (Related) was performed for each input mode taking mean call duration as the dependent variable with the low and high spatial groupings being used as the categorical variable. Once again, significant differences were observed between high and low spatial ability groups on the connected word interface, {F(1,82) = 6.746, $p < 0.05$}, and the keypad interface; {F(1,82) = 17.002, $p < 0.001$}. No significant differences were observed between low and high spatial ability subjects on call duration for the isolated word interface.

Individual Difference #2

Mean call durations are significantly longer for low spatial ability users than for high spatial ability users when using connected word or keypad input to complete a complex task.

The individual differences in cognitive ability were also analysed in relation to the number of errors that subjects made when using each of the three interfaces. No significant differences were observed between low and high spatial and verbal ability subjects in terms of the number of errors they made when interacting with any of the three data entry interfaces. Table 4.7 shows the mean error counts for subjects in terms of spatial and verbal ability across the three input modes.

Cognitive Ability	Isolated Word	Connected Word	Keypad
low verbal	0.76	0.48	0.36
high verbal	0.48	0.19	0.19
low spatial	0.84	0.54	0.37
high spatial	0.39	0.12	0.17

Table 4.7 Mean Error Scores for Cognitive Ability Across Input Modes

The results obtained from the analysis of subjects' cognitive abilities indicate that verbal ability and spatial ability are important in determining call duration differences on a task of this kind, for both connected word and keypad input. What is interesting to note here is the significant result for verbal ability differences on call duration performance. In most of the research that has been reported in the area of individual differences in computer-based tasks, verbal ability has been demonstrated not to have a significant effect on the subjects' performance of the task. This is perhaps not surprising given the fact that most of these results were based on subjects using graphical user interfaces where speech (and therefore verbal ability) does not play an integral part in the interaction. Verbal comprehension takes on a more significant role because information can only be conveyed to the user by spoken messages (and tone prompts, to a lesser extent). This is in contrast to screen-based interfaces which can provide information via written text, icons, cursor display and graphics. Therefore the more proficient the individual is in his or her inductive reasoning ability, as measured by the verbal ability sub-scale of the AH4 Group Test of General Intelligence, the better the expected performance.

Spatial ability plays a significant role in call duration performance with keypad and connected word interface. Subjects who have high spatial ability perform better using the keypad and connected word interfaces than low spatial ability subjects, possibly because their spatial ability allows them to develop a more efficient mental model of the task and the system which in turn allows them to navigate through the system in a more efficient manner than their low spatial ability counterparts.

There were no significant cognitive differences found for performance (call duration and error rates) using the isolated word interface. This could be due to the constrained dialogue interaction used in this type of interface, which decreases the opportunity for individual variability in performance. There may, however, be differences in how low and high cognitive ability subjects perceived the usability of the isolated word interface. This possibility is explored in the next section which analyses subjects' responses to the Likert attitude and usability questionnaire.

4.3.3. Individual Differences in Cognition - Attitude Analysis

The Likert attitude questionnaires provide valuable data on a 7-point scale of subjects' attitudes with respect to data entry mode. Table 4.8 shows the mean attitudes to connected word, keypad and isolated word input modes for low and high verbal ability.

Input Mode	Low Verbal Ability	High Verbal Ability
Isolated Word	5.2	4.9
Connected Word	5.6	5.3
Keypad	5.6	5.6

Table 4.8 Mean Attitude Scores for Verbal Ability

A One-Way ANOVA (Related) was performed for each input mode taking mean attitude as the dependent variable and the level of verbal ability as the categorical variable. No significant effect was found for verbal ability on attitude for any of the input modes, although connected word differences were close to significance, {F(1,82) = 3.789, p < 0.055}.

A similar type of analysis was carried out for spatial ability. Table 4.9 shows the mean attitudes to connected word, keypad and isolated word input modes for low and high spatial ability.

Input Mode	Low Spatial Ability	High Spatial Ability
Isolated Word	5.3	4.8
Connected Word	5.7	5.1
Keypad	5.6	5.5

Table 4.9 Mean Attitude Scores for Spatial Ability

Once again, a One-Way ANOVA (Related) was performed on the data for each input mode taking mean attitude as the dependent variable and level of spatial ability as the categorical variable. A significant difference in attitude was found between low spatial and high spatial groups for connected word data entry, {F(1,82) = 10.665, p < 0.002}.

Individual Difference #3

Low spatial ability users have a significantly more positive attitude towards connected speech input mode than high spatial ability users.

Further analysis was carried out comparing attitudes to input mode for high and low verbal ability and high and low spatial ability separately. For subjects with low spatial and low verbal ability, Related T-Tests showed that there was no significant difference between keypad and connected word data entry mode (p=0.495 for low

verbal and p=0.65 for low spatial). There were, however, highly significant differences between keypad and isolated word entry, {t = 3.005, df = 39, p < 0.01} for low spatial and {t = 2.835, df = 41, p < 0.01} for low verbal. Significant differences were also found between connected word and isolated word entry,
{t = 5.445, df = 39, p < 0.001} for low spatial and {t = 4.349, df = 41, p < 0.001} for low verbal ability subjects.

These differences can be related to the amount of effort required by the low cognitive ability subjects to use the isolated word input version of the automated catalogue service in comparison to both the connected word and keypad input modes. Evidence comes from their scores on specific attitude statements. For example, low verbal and low spatial ability subjects found the isolated word input mode more confusing to use than either of the other two input modes; this group of subjects also felt more out of control when using the isolated word version of the service. In addition, the low verbal ability subjects felt the isolated word input mode was less efficient to use than the connected word and keypad input modes.

Individual Difference #4

Low cognitive ability users have a significantly more positive attitude towards connected word and keypad input modes in comparison to isolated word input mode.

For subjects with high verbal and high spatial ability, Related T-Tests showed that there were significant differences between keypad and isolated word entry,
{t = 5.165, df = 43, p < 0.001} for high spatial ability and {t = 5.450, df = 41, p < 0.001} for high verbal ability. Significant differences were also found between connected

word and isolated word entry, {t = 3.651, df = 43, p < 0.001}, for high spatial ability and {t = 4.293, df = 41, p < 0.001} for high verbal ability. In addition, significant differences were also found between keypad and connected word entry, {t = 3.312, df = 43, p < 0.01} for high spatial ability and {t = 3.679, df = 41, p < 0.001} for high verbal ability.

High cognitive ability subjects felt more in control of the interaction when they were using the keypad input version of the automated catalogue service in comparison to the connected word and isolated word input versions. This group of subjects also enjoyed using the keypad version of the service more than the other two versions. The high cognitive ability subjects also expressed feelings of being more flustered when using either of the speech input modes, this could be due to the fact that they felt more constrained when using these versions of the service in comparison to how they felt when using the keypad version of the automated catalogue service.

Individual Difference #5

High cognitive ability users have a significantly more positive attitude towards connected word and keypad input modes in comparison to isolated word input mode and have a more positive attitude towards keypad input than connected word as an input mode.

A number of conclusions may be drawn from these results. Firstly, all subjects, whether of low or high cognitive skills (i.e. spatial and verbal ability), rate isolated word entry significantly worse than either keypad or connected word entry. Secondly, subjects high in cognitive skills significantly prefer keypad over connected word as a mode of entry. These can be attributed, in part, to specific aspects of usability, such as the perceived efficiency of the three services and, to the different amount of effort required to use the three input modes. In addition, an examination of order of presentation has shown that as users cease to be novices the isolated word input mode is rated significantly worse than either connected word or keypad input in terms of perceived usability.

4.3.4. Individual Differences in Personality - Objective Measures Analysis

The scores from the NEO-PI-R personality inventory were converted into two groupings - low and high for each of the five traits in terms of subjects scores on the sample distribution. Table 4.10 shows the mean call duration for use of the three interfaces for each of the five personality traits - Neuroticism, Extroversion, Openness, Agreeableness, Conscientiousness -for subjects in the low and high personality trait groupings.

	N		E		O		A		C	
	Low	High	Low	High	Low	High	Low	High	Low	High
IW	170	174	170	175	171	174	173	172	170	174
CW	158	156	158	154	158	155	156	156	160	152
KP	114	114	115	112	114	114	113	114	110	117

Table 4.10 Mean Call Duration (in secs) for Subject Scores on the NEO-PI-R

A One-Way ANOVA (Related) was performed taking the mean call duration for each interface as the dependent variable and each of the five traits as the categorical variable. No significant differences were found between the low and high trait groupings for any of the five factors in relation to the three interfaces.

A similar type of analysis was carried out for error rates. Once again no significant differences were observed between the high and low groupings for each of the five personality factors and performance on the three interfaces. This finding is not surprising given the fact that there were so few errors made by subjects when they were using connected word, isolated word and keypad interfaces.

Personality is regarded as being one of an individual's characteristics which remains stable across different tasks and different situations (van Muylwijk, van der Veer, Waern, 1983). Most of the previous work which examines the effects of personality on task performance has been in the areas of text editing and information searches. The results from this earlier work have shown (Koubek, Le Bould and Salvendy, 1985) that no effects for personality on performance have been observed. The results obtained from the analysis of personality in this first experiment therefore concurs with the findings reported in these other domains.

4.3.5. Individual Differences in Personality - Attitude Analysis

Differences in attitude towards the three input modes due to personality traits were examined next. A One-Way ANOVA (Related) was performed taking the mean attitude score to each interface as the dependent variable and each of the five traits as the categorical variable. No significant differences were observed between the high and low trait groups for any of the three interfaces. This result is consistent with the results of the analysis which investigated the relationship between personality and user performance when using connected word, isolated word and keypad interfaces.

4.3.6. Individual Differences in Age and Gender - Objective Measures Analysis

The effects of age differences on subjects performance on the three interfaces were explored by carrying out a One-Way ANOVA (Related) on the experimental data. In order to carry out the analysis the subjects were divided into three age categories: 18-38, 39-59 and 60+. Figure 4.2 shows the mean call duration scores for the three age groups in relation to the three interfaces.

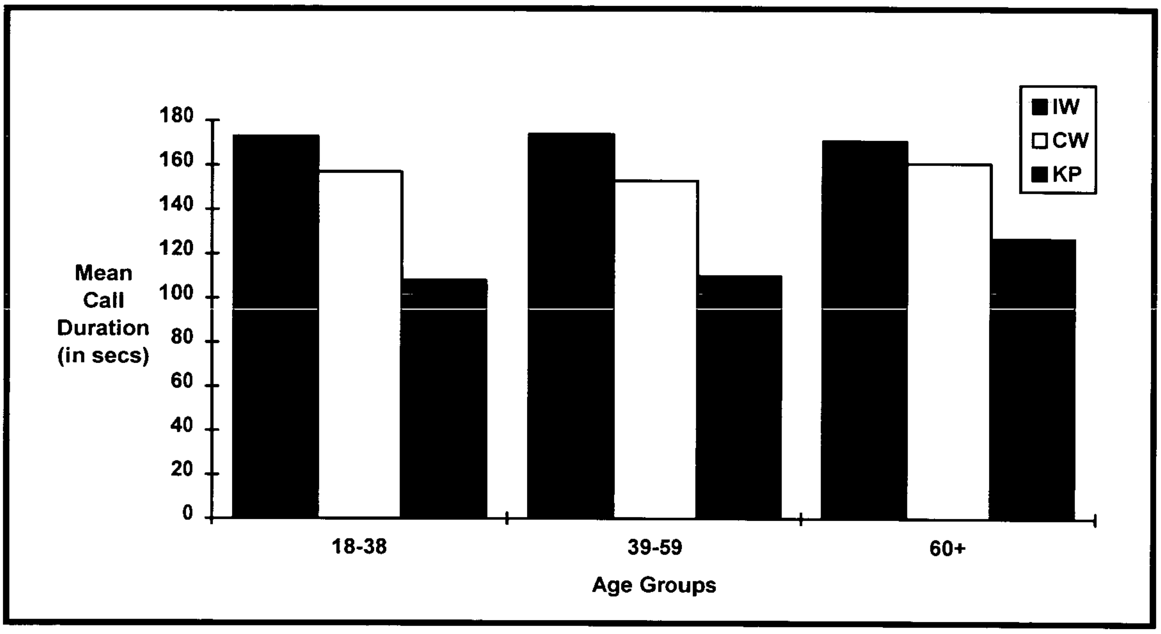


Figure 4.2 Mean Call Duration for Age between Input Modes

A significant difference was observed between the three age groups on the keypad interface, {F(2,81) = 10.896, p < 0.001}. No significant differences were found between the age groups on the connected word and isolated word interface in terms of call duration performance. Once again there was no significant differences observed for age on any of the three interfaces in relation to the number of errors made. It has already been stated that people in general made only one or two errors.

Individual Difference # 6

Older subjects take significantly longer than younger subjects to complete a task when using keypad input.

Gender differences in call duration performance for the three types of interfaces were then explored by a One-Way ANOVA (Related). Figure 4.3 shows the mean call duration times for connected word, isolated word and keypad input modes for males and females.

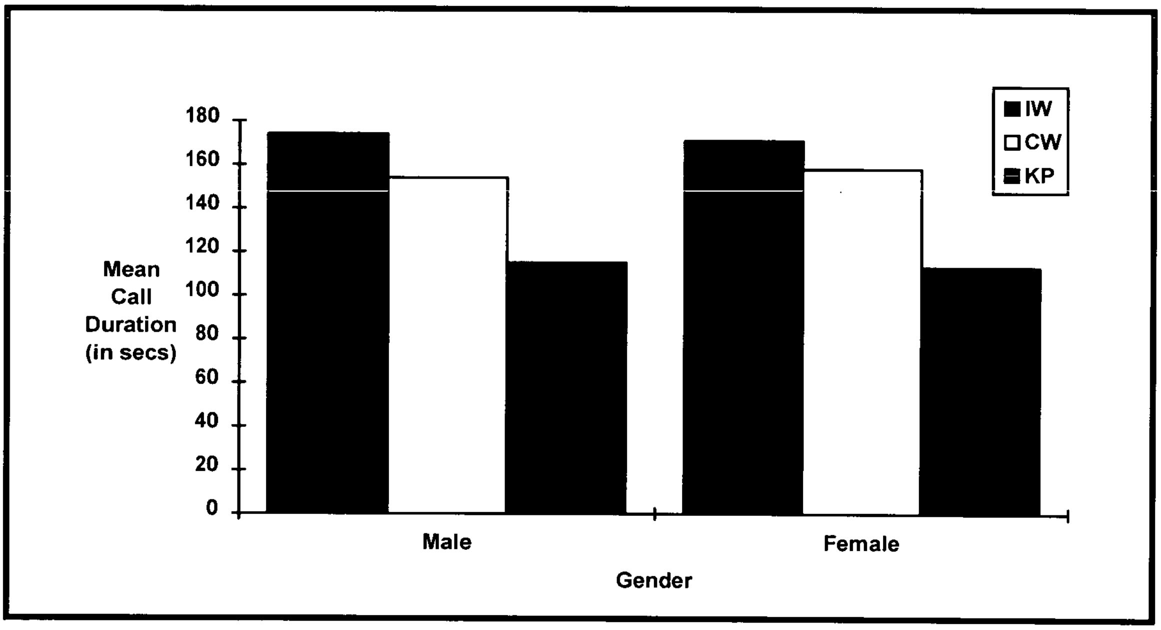


Figure 4.3 Mean Call Duration for Males and Females

No significant differences were observed between males and females on the connected word interface, isolated word interface or the keypad interface. The analysis was then repeated for a second time taking number of errors made as the measure of performance. This analysis also produced no significant findings and is consistent with the other results obtained in this experiment in relation to individual characteristics and performance when error rates are taken as the measure of performance.

The results from the experiment reported here indicate that in terms of task behaviour as measured by call duration length and number of errors made, no significant differences exist between the two sexes when using any of the three interfaces. There could however be differences in their attitudes to the three interfaces (this possibility will be considered in the next section). Another reason for the results found here could lie with the nature of the task itself.

A significant difference was found for age on the keypad interface when call duration was taken as a measure of performance. This could be due to the fact that younger subjects in the experiment were more aware of new technology (or were perhaps more dextrous) and were able to use the telephone keypad more like a computer keyboard, something with which the older subjects may not have had as much experience. This could be investigated further by looking at the effects of age on subjects' perception of the usability of the keypad interface as measured by the Likert usability metric.

4.3.7. Individual Differences Age and Gender - Attitude Analysis

Gender differences in attitude to the three types of data input modes were explored by One-Way ANOVAs (Related) and Related T-Tests. Table 4.11 shows the mean attitude scores for connected word, keypad and isolated word input modes for males and females.

Gender	CW	KP	IW
Male	5.4	5.6	5.1
Female	5.5	5.6	5.0

Table 4.11 Mean Attitude Scores for Gender

Two sets of significance tests were performed on these data. In the first instance a series of One-Way ANOVAs (Related) were performed for each input mode taking mean attitude as the dependent variable and sex as the categorical variable. No significant difference was found for gender within each input mode. As can be seen from Table 4.11, the means for each input mode across gender are very similar.

The data was also analysed by conducting a series of Related T-Tests. This analysis found that males significantly preferred connected word entry to isolated word entry mode, $\{t = 3.774, df = 37, p < 0.001\}$, as did females, $\{t = 4.683, df = 45, p < 0.001\}$. There was also a significant preference for keypad entry over isolated word entry for males, $\{t = 4.386, df = 37, p < 0.001\}$ and a similar preference was found for females, $\{t = 3.828, df = 45, p < 0.001\}$. No significant difference was found when attitudes towards connected word and keypad were compared for both sexes.

Individual Difference # 7

No significant differences in attitude were found between males and females for any of the three input modes. However, both male and female users have a significantly more positive attitude towards connected word and keypad input modes than towards isolated word input mode.

The effect of age on differences in attitude to the three types of data input modes was explored in a similar manner to the gender difference analysis, by conducting a One-Way ANOVA (Related) and Related T-Tests. Table 4.12 shows the mean attitude scores to the three input modes for each age group.

Age Group	CW	KP	IW
ages 18-38	5.4	5.7	5.0
ages 39-59	5.3	5.4	4.8
age 60+	5.6	5.5	5.3

Table 4.12 Mean Attitude Scores for Age

A One-Way ANOVA (Related) was performed for each of the input modes taking mean attitude as the dependent variable and age group as the categorical variable. No significant difference was found overall between the three age groups in attitude for any of the input modes.

The results of this analysis indicated that the 18-38 age group significantly preferred connected word to isolated word entry, {t = 3.723, df = 31, p < 0.001}; significantly preferred keypad entry to isolated word entry, {t = 5.754, df = 31, p < 0.001}; and also preferred keypad mode of entry to connected word,

{t = 2.947, df = 31, p < 0.01}. The 39-59 age group also significantly preferred connected word entry mode to isolated word entry, {t = 3.901, df = 29, p < 0.00}; and significantly preferred keypad entry over isolated word entry, {t = 3.542, df = 29, p < 0.001}. The analysis of the data from the 60+ age group produced only one significant finding with the subjects preferring connected word entry mode to isolated word entry, {t = 2.643, df = 21, p < 0.01}.

Individual Difference # 8

Younger users significantly prefer keypad input to either connected word or isolated word input modes whereas older users prefer connected word input.

4.3.8. Multiple Regression Models

Multiple regression is a statistical technique which allows scores on one variable or measure (known as the *criterion measure*) to be predicted, to a certain extent, by scores on two or more variables or measures (known as *predictor variables*). The aim of the regression equation is to produce a 'line of best fit' which offers a precise estimate of the relationship between the criterion measure and predictor variables. The extent to which each predictor variable predicts the value of the criterion variable is given by the *regression coefficient.* This value is the result of a correlation between the criterion variable and each of the predictor variables. A multiple regression also produces a *multiple regression coefficient* (R^2) which gives an estimate of the amount of variance in the criterion measure that has been 'explained' by a combination of the predictor variables.

On the basis of the results obtained from this experiment a multiple regression analysis was carried out, for keypad and speech input modes, taking call duration as the criterion variable and age, verbal ability and spatial ability as the predictor variables. A *forward stepwise regression* technique was used. This form of regression analysis enters the most significant predictor variable at the first step and continues to add or delete variables until none can significantly improve the 'fit'; providing a model of call duration performance for example, based on a significant subset of predictor variables.

The results of the regression analysis are shown in Table 4.13. Standardised coefficients are given for the predictor variables and R^2 denotes the proportion of variance accounted for in the criterion variable by the predictor variables.

Measure	Age	Verbal Ability	R^2
Call Duration (KP)	.417**	-.265***	.435**
Call Duration (CW)		-.392*	.143*

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4.13 Significant Predictors of Call Duration Performance for Keypad and Connected Word Input mode

The results shown in Table 4.13 indicate that age and verbal ability can be used to predict call duration performance when the mode of input is keypad whereas for connected word input, verbal ability alone can be used to significantly predict call duration performance. The pattern of regression coefficients obtained in Table 4.13 has a fairly straightforward interpretation when looked at in conjunction with the other results obtained in this experiment. Age can be said to be a predictor of call duration for the keypad interface because older subjects may not be as familiar as others when it comes to using new technology or as was the case here, using the keypad on the telephone in a manner which is similar to a keyboard on the computer. People with high verbal ability may take less time to complete the task because they can process and act upon the serial information they are receiving over the telephone more quickly than those subjects with low verbal ability. This effect for verbal ability also applies in the case of the connected word version of the automated catalogue service.

4.4. Conclusions

The aim of this experiment was to examine, or "assay" in the Egan and Gomez terminology, which individual characteristics are important in determining a subject's performance on an automated catalogue ordering task. The subjects used three functionally similar interfaces to the automated catalogue service. The most significant differences in performance were observed when subjects were using the keypad interface.

The results from this experiment indicate that cognitive abilities, especially verbal ability, and age contribute significantly to the variation observed in subjects' performance and to a lesser extent their perception of the usability of the systems, especially in relation to the keypad interface. As stated in Chapter 1 of this thesis, Shneiderman (1987) suggested the following measures as being suitable for quantifying the efficiency and usability of an interactive system - speed of performance, rate of error, subjective satisfaction, time to learn and retention over time. The results discussed for this experiment have relied on three measures similar to the ones put forward by Shneiderman - call duration, number of errors and attitude towards the automated telephone service. However by considering the way in which these results are affected by individual characteristics such as age and verbal ability, an explanation of these performance measures is available. Therefore these individual characteristics can be used as part of a process to improve the usability of an interface for an automated telephone service.

Overall, the results obtained from this experiment indicate that in terms of assaying individual differences, the salient individual characteristics for keypad and connected word performance (measured by call duration) on the automated catalogue service are age, verbal ability and spatial ability. No significant results were obtained for subjects' performance on the isolated word version of the automated catalogue service. When the relationship between these salient characteristics and subjects' performance on the different input modes was explored by multiple regression analysis, it was found that age and verbal ability could be used to predict call duration performance on the keypad version of the automated catalogue service, whereas verbal ability could be used to predict call duration performance on the connected word version of the service. It can be said therefore that the experimental aims set out at the beginning of this chapter have been met.

Chapter 5: Isolating Individual Differences

5.1. Introduction

The results outlined in Chapter Four demonstrated that subjects' performance (measured by call duration) when they interact with spoken language dialogue systems can be predicted on the basis of specific individual characteristics such as age and verbal ability. The overall aim of the experiment reported in this present Chapter is to investigate the second step of the Egan and Gomez methodology outlined in Chapter Three: *isolating*. This involves breaking down the task that subjects are required to carry out into smaller components. The purpose of this process is to isolate where the salient individual differences identified in the previous experiment have their greatest effect on subjects' performance. Once this information has been obtained, the final step is to look for ways to accommodate for these individual differences through dialogue design, training or performance support.

In order to investigate this, a more complex automated telephone service was designed which was based on an audio on demand music catalogue using extensive menu selection. The automated music catalogue uses a hierarchically-structured database with menus from three to five items in length. The organisation of the database is designed to reflect the layout of a typical music store in which CD's are grouped into categories (i.e. types of music), artists and albums:

- the *category menu*offers five categories of music available in the catalogue: Blues, Jazz, Classical, Rock & Pop and Folk

- the *artist menus* offer four artists, musicians or composers for each music category :

  Blues - John Lee Hooker,Howlin' Wolf, EricClapton and Muddy Waters
  Jazz - John Coltrane, Chet Baker, StanGetz and Art Pepper
  Classical -Bach, Mozart, Beethoven andBrahms
  Rock & Pop - TheBeatles, The Rolling Stones, Sting, REM and Neil Young
  Folk -Joni Mitchell, Bob Dylan, TheCorries and Steeleye Span

- the album menus offer three albums for each artist or composer:

e.g. John Coltrane - A Love Supreme, Blue Train and Private Recordings & Curios

- the track menus offer three tracks for each album:

e.g. A Love Supreme - Acknowledgement, Resolution and Pursuance

The scenario for this experiment is that users can interact with an on-line music catalogue via a telephone keypad to compile a 'personalised' CD. As with the keypad version of the automated catalogue service used in the previous experiment, user's inputs (key presses) when interacting with the automated music catalogue are handled automatically by the telephone interface card, which passes the input on to the simulation software. The role of the operator in this dialogue service is to monitor the interaction and make note of any irregularities. The experimental task requires subjects to order four tracks: two of which have been specifically identified to be selected while for the other two tracks subjects are required to listen to a choice of three tracks before making a selection. The service provides 'audio on demand' at the level of individual tracks which allows subjects the opportunity to listen to up to 30 seconds of each track available in the automated music catalogue. Subjects attended the laboratory four times at one week intervals and each time they attended they were given a different list of categories from which to make their selection.

The keypad input service begins with a message welcoming users to the automated music catalogue. Subjects are then presented with the category menu offering the categories of music available in the automated music catalogue (e.g. 'For Blues press 1, for Jazz press 2, for Classical press 3, for Rock & Pop press 4, for Folk press 5'). After this menu has been presented subjects are informed that they can request more information from an options menu (which, they are informed, is available at any point in their interaction with the automated music catalogue) by pressing the star key (*) on the telephone keypad. The options menu provides users with functions such as the opportunity to leave the music catalogue and to hear a menu repeated again. In addition, it also allows users to return to the category menu (top level) of the automated music catalogue, regardless of where they are in the dialogue structure. The automated music catalogue also offers a menu interrupt facility which allows users to interrupt a menu prompt before it finishes, if they know what key they wish to press. If the subject inputs an illegal key (e.g. pressing the square key '#' instead of the star '*' key) the system will respond by asking the subject to input the information again.

After selecting a music category, subjects have to select a particular artist or composer from the artist menu. For example, if a subject chooses the classical music category the automated music catalogue prompts the subject to select a composer (e.g. 'There are four classical composers available, they are Bach, Mozart, Beethoven and Brahms. For Bach press 1, for Mozart press 2, for Beethoven press 3, for Brahms press 4'). If the subject selects Beethoven, they would be informed that there are three albums available by Beethoven in the automated music catalogue and prompted to make an album selection (e.g. 'Three albums are available by Beethoven, they are Piano Sonata No 8, Piano Concerto No 5 and Symphony No 5. For Piano Sonata No 8 press 1, for Piano Concerto No 5 press 2, for Symphony No 5 press 3'). Once a selection has been made, subjects have successfully navigated down to the track menu of the automated music catalogue. In the track menu subjects are presented with three options to choose from. For example, if a subject has chosen Piano Sonata No 8 at the album level they are offered the following three tracks to choose from - Grave, AdagioCantabile and Rondo (Allegro).

Once a track selection has been made subjects are presented with another menu from which to make a selection (this is referred to as menu 1 in the automated music catalogue design protocol). This menu gives the subject an opportunity to play the track that has been selected (and hear a 30 second extract from it), add it to their CD without listening to it or make a different selection. If a subject chooses the 'different selection' option they are taken back to the tracks menu and played the menu options available at that point. If the subject chooses to listen to the track, they are presented with another list of options after the service has played the extract from the track (this is referred to as menu 2 in the automated music catalogue design protocol). Subjects have the opportunity to play the track again, add the track to their CD or make a 'different selection'. Once again the different selection option will take subjects back to the track level menu. If the subject chooses to add the track they are presented with another menu selection (referred to as menu 3 in the automated music catalogue design protocol). This menu is essentially a navigation menu with users having options which include the ability to move to another artist within the music category subjects are currently in and to move to other music categories. After subjects have added a fourth track to their CD the system informs that their CD is now complete and plays back the list of tracks which have been ordered, including the name of the artist or composer. No confirmation or rejection is required from users at this point. The service concludes the interaction by thanking subjects for using the automated music catalogue.

Two objective measures of performance were obtained - *interrupt rate* and *silence rate*. Interrupt rate can be defined as the number of times a subject keyed in a response before the end of a system prompt divided by the overall number of responses made by a subject to menu prompts during an experimental trial. 'Silence' was recorded when a subject did not respond by the end of a time-out period following a system prompt (5 seconds). A subject's silence rate was calculated in a similar fashion to their interrupt rate. Unlike the previous experiment in Chapter Four, call duration was not used as an objective measure of performance in the experiment reported in this chapter. This was due to the fact that it would not provide a reliable measure of performance because of subjects' ability to 'interrupt' system prompts and music samples. Subjective data was obtained from subjects responses to the Likert attitude questionnaire.

There were four main aims in this experiment. First, to isolate those parts of the interaction where the salient individual characteristics have their greatest effect. Secondly, to reproduce for a more complex task, the "assay" found in the last experiment between user performance and age and verbal ability. The third aim was to investigate further characteristics of individuals which may affect their interaction with a spoken dialogue system. In the case of the experiment described here the impact of information processing ability and short-term memory capacity were assessed in addition to those individual abilities tested in the first experiment. These were chosen because of their possibility of affecting performance on a task which required users to make selections from menu lists they hear and also to navigate, relatively quickly (due to system time out), their way through the hierarchical database on information prompts provided by the system and finally to explore the impact of learning on subjects' performance.

5.2. Method

5.2.1. Design

The experiment employed a repeated measures design, with subjects visiting the laboratory four times at one week intervals. Each time the subjects attended they were given a different list of music categories from which to make their selection.

5.2.2. Subjects

In all, 30 subjects took part in this experiment none of whom had previously used spoken language dialogue systems. The distribution of gender and age group is shown in Table 5.1.

	Ages 18-38	Ages 39-59	Age 60+	Total
Male	10	2	3	15
Female	6	7	2	15
Total	16	9	5	30

Table 5.1. Age and Gender Breakdown of Subjects

5.2.3. Materials

Four psychometric tests were used in the experiment:

NEO - Personality Inventory Revised (McCrae & Costa, 1987). This test provides a comprehensive measure of personality using a five-factor approach which is widely used in trait psychology.

AH4 Test of General Intelligence (Heim, 1970). This test is widely used to assess general ability including verbal, numerical and spatial reasoning skills.

The Paced Auditory Serial Addition Task (PASAT) (Gronwall, 1977). This test provides a measure of information processing capacity. This test requires subjects to listen to a series of digits (1 to 9) played on a tape recorder. The subject must add the numbers in the following way: add the first number to the second number and tell the tester the answer, add the second number to the third number and tell the answer, and so on. Usually, 61 digits are played (which gives 60 answers). The task can be made more difficult by reducing the time interval between digits. One common form of the test allows the subject to practise on a set of 10 digits and then presents them with a series of digits at four second intervals followed by a series of digits at two second intervals. This was the form adopted for use in the experiment reported in this chapter.

The Rey Auditory Verbal Learning Test (AVLT) (Lezack,1983). This test assesses both short-term memory and long-term memory. In this test subjects are given five trials to learn a list of 15 words which they are asked to recall: (a) immediately after each trial, which produce scores called AVLT 1-5, (b) after recalling words presented as an interference list which is presented subjects after their responses on the fifth trial have been recorded and where memory score for the interference list (made up of 15 different words) is called AVLT B and subjects' memory score for the original list after recalling the words on the interference list is called AVLT 6.

A Likert attitude questionnaire was also used in this experiment. Each subject received the same questionnaire, with the order of the Likert statements randomised on each of the four experimental trials. In addition, an instruction sheet explaining how to use the automated music catalogue was available on each trial. A different list of items to order was provided for each trial.

5.2.4. Apparatus

For the purposes of the experiment reported here, an IBM-compatible 486 PC was set up in an office near the laboratory where the subject sat to provide a connection to the telephone line and allow keypad input from the subject at the other end of the telephone. A telephone was also provided in the laboratory for the subjects to use as part of the experiment.

5.2.5. Procedure

For the purposes of carrying out this experiment subjects attended the laboratory at four one week intervals. On their first visit, subjects completed the NEO-PI-R personality inventory. After a short break the subject was given a sheet of instructions describing how to use the automated music catalogue and a list of the music tasks they were required to order. After completing the task, each subject completed a Likert attitude questionnaire. The session finished with a semi-structured interview where subjects were asked specific questions such as "Was there anything you wanted to do and did not know how to do?" and "Was there anything that happened that you felt should not have happened?". Subjects were also asked to comment on any aspect of the experience they had just gone through.

On their second visit subjects completed the AH4 Test of General Intelligence when they first arrived. After a short break they were given the instruction sheet describing how to use the automated music catalogue together with a new list of items to order. After completing the task, subjects completed the Likert attitude questionnaire. Finally asemi-structured interview was carried out.

On their third visit subjects completed the Paced Serial Addition Task (PASAT) and after a short break completed the Rey Auditory Verbal Learning Test (AVLT). After a short break subjects were given the same instruction sheet once again along with a new list of items to order. After completing the task subjects were given the Likert attitude questionnaire and the session finished off with the semi-structured interview.

On the subject's fourth visit - psychometric tests were not required. Subjects were given the same instruction sheet on how to use the automated music catalogue together with a new list of items to order. After completing the task, subjects were given the Likert attitude questionnaire and this was followed by the semi-structured interview.

5.3. Results

In each experimental trial, the following objective measures of subject performance were made:

- task completion rates
- menu interrupt rate
- silence rate per subject

5.3.1. Task Completion Rates

The performance of the subjects in terms of successful task completion across the four trials is given in Table 5.2. Entries in columns represent actual numbers of subjects.

Outcome	Trial 1	Trial 2	Trial 3	Trial 4
1.Completed task successfully	10	15	16	20
2.Did not follow instructions but completed task	12	8	11	8
3.Did not order correct four tracks	2	5	2	1
4. Failed	6	2	1	1

Table 5.2. Subject Task Completion Performance

Outcome 1 refers to those subjects who followed the instructions exactly and completed all tasks successfully. Outcome 2 refers to those subjects who did not follow the instructions but still managed to order the correct four tracks from the correct albums. Subjects in this category may, for example, have chosen a CD track as required without listening to all the tracks available (as instructed). Outcome 3 refers to those subjects who did not follow the instructions and did not order four tracks from the correct albums. They did, however, order four tracks in all. Outcome 4 refers to those subjects who failed outright. At some point, these subjects made three consecutive errors (e.g. illegal key press, failure to respond to a system prompt) which resulted in the termination of the dialogue ("I'm sorry there seems to be a problem. Please hang up").

A binomial sign test was performed on these data. In order to carry out an analysis looking for learning effects the subjects were split into two groups - those who successfully completed the task (outcome 1) and those who did not (outcome 2, outcome 3, outcome 4). The results from this analysis showed that there was a significant improvement in the number of subjects successfully completing the task between trial 1 and trial 4, $p < 0.01$. This result shows that a degree of learning appears to have taken place over the four trials.

5.3.2. Interrupt Rate and Silences

Table 5.3 shows, for each experimental trial, the mean of subjects' total responses during the trial, the mean percentage of interrupt attempts and the mean percentage of user silences. As stated earlier, an 'interrupt attempt' was taken to occur when the subject keyed in a response before the end of a system prompt; 'silence' was recorded when a subject did not respond by the end of the time-out period following the prompt (5 seconds). The percentage interrupt rate per subject is defined as the number of interrupt attempts during the experimental trial divided by the overall number of responses to menu prompts during the trial expressed as a percentage. The percentage silence rate is defined similarly.

	Trial 1	Trial 2	Trial 3	Trial 4
Mean user responses	36.9%	39.4%	42%	41.4%
Mean interrupt rate	51%	73%	81%	82%
Mean silence rate	6%	3%	2%	1%

Table 5.3. Distribution of Subject Dialogue Events across Trials

In order to test for learning effects between trial 1 trial 4 a binomial sign test was performed on each of the dialogue event measures. A significant difference was found between the mean percentage interrupt rates in trial 1 and trial 4, $p < 0.001$. A significant difference was also observed between the mean percentage silence rates in trial 1 and trial 4, $p < 0.05$. No significant difference was found for overall subject responses to menu prompts.

The fact that the interrupt rate increased significantly gives an indication of the learning that took place as the subjects progressed through the four experimental trials. As subjects became more familiar with the service they knew they could progress through the system without having to listen to the whole of a system prompt before deciding what to do next. Similarly, the fact that the number of silences decreased significantly over the four trials indicates that subjects became more familiar with the service through use.

5.3.3. Attitude Scores

As indicated earlier, following each use of the service, subjects completed a Likert attitude questionnaire consisting of 32 attitude statements each with a 7-point Likert response scale with a mid-neutral point. The Likert attitude questionnaire provided a measure of subjects' attitudes towards the automated music catalogue. Figure 5.1 shows the mean of subjects' attitudes across the four experimental trials.

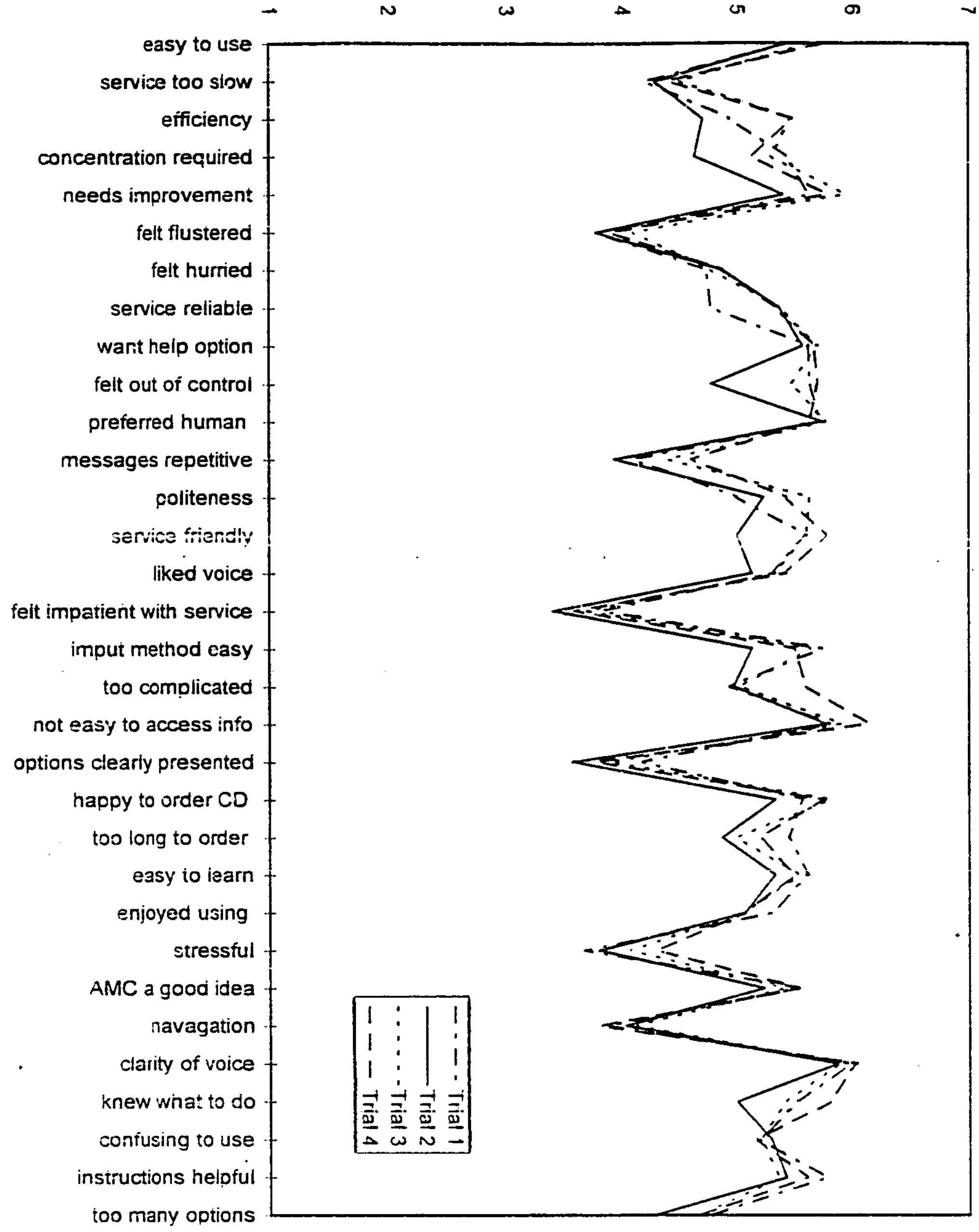


Figure 5.1
Mean Subject Attitude Scores Across Experimental Trials

A One-Way repeated measures ANOVA was performed upon the attitude data. No significant difference was found in subjects' attitude towards the automated music catalogue across the four experimental trials. This result seems surprising as the attitude to the service might have been expected to improve over the four trials.

especially as the overall performance measures showed an increase in interrupt rate and a decrease in the silence rate over the trials.

5.3.4. Age Differences

The effect of age on the number of times subjects interrupted a system prompt was the first analysis undertaken. Figure 5.2 shows the mean number of interrupts per age group across the four experimental trials.

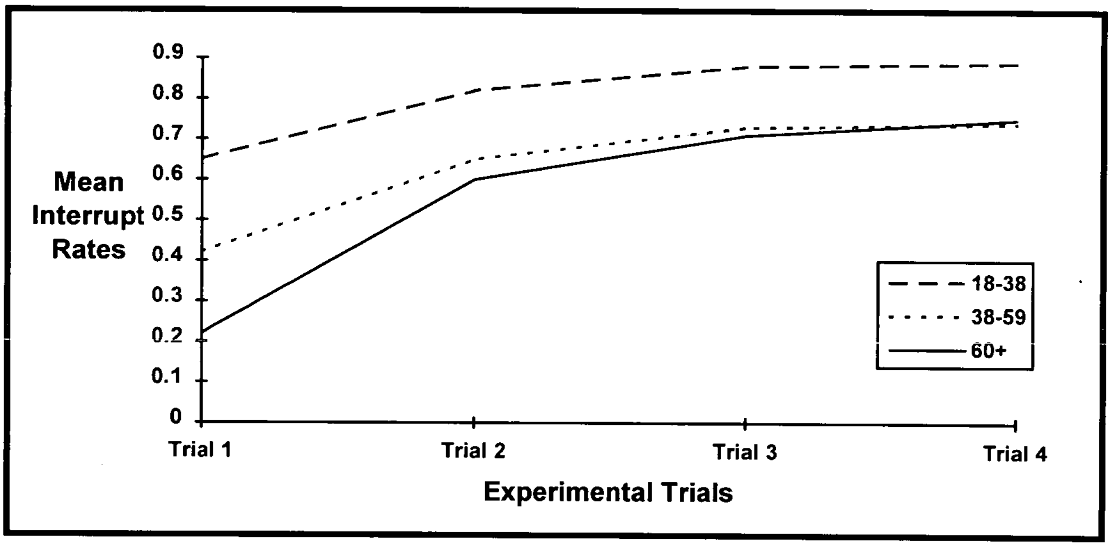


Figure 5.2 Mean Rate of Interrupt for Age

A One-Way repeated measures ANOVA was performed on this data. A significant difference in interrupt rate was found between the age groups across the four experimental trials, {F(2, 27) = 11.14, p < 0.001}. Related t-tests over trials 1 and 4 showed a significant increase in the use of interrupt by the 18-38 age group {t = -6.016, df = 15, p < 0.01} and a highly significant increase for the two older age groups {38-59: t = -5.240, df = 8, p <0.01; 60+: t = -4.901, df = 4, p < 0.001}. A Tukey HSD test showed that for trial 1 there was a highly significant difference in mean interrupt rate between the 18-38 and 60+ age groups (p < 0.001) and a significant difference between the 18-38 and 39-59 age groups (p < 0.02). The difference between the 38-59 and 60+ age groups was not significant. With respect to trial 4, there were significant differences between the 18-38 and 60+ age groups (p < 0.05) and between the 18-38 and 39-59 age groups (p < 0.01). These results show that it was subjects in the oldest age group (60+) who improved the most over the four trials but their performance (as measured by interrupt rates) was still significantly worse from the youngest age group even at the end of the four experimental trials.

Individual Difference # 9

The interrupt rates for older users are still significantly lower than for younger users even after four uses of a service.

The number of silences that occurred during the subjects' interactions was then analysed. Figure 5.3 shows the mean number of silences per age group across the four experimental conditions.

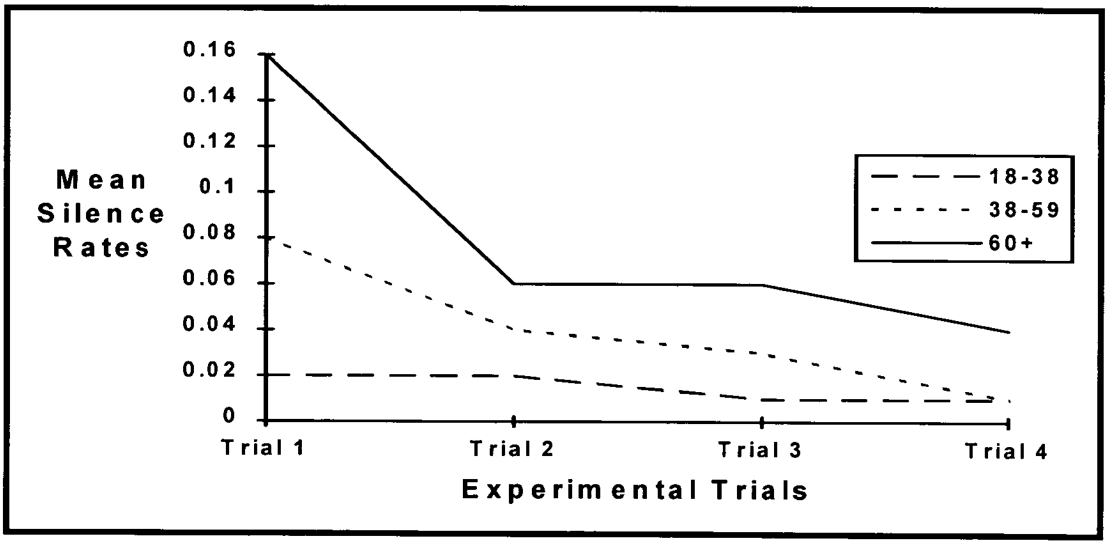


Figure 5.3: Mean Rate of Silence for Age

A significant difference was found overall for age across the four experimental trials, {F(2,27) = 4.03, p < 0.05}. Related t-tests over trials 1 and 4 showed no significant differences for any of the age groups. Similarly a Tukey HSD test showed no significant within-trial differences in silence rate between any of the age groups for trials 1 and 4.

Individual Difference # 10

Both younger and older users significantly improve upon their silence rates with keypad input after four uses of a service.

When the Likert attitude data were analysed no differences in attitude were observed between the three age groups across the four experimental trials. This reaffirmed the findings of the first experiment which showed that there was no difference in attitude between the three age groups towards the keypad version of the automated music service. This is similar to the result obtained in the previous experiment when attitudes towards the keypad version of the automated catalogue service were analysed in terms of age differences.

The results obtained from the related t-tests and the Tukey HSD demonstrate a more rapid increase in interrupt behaviour by the 60+ age group when compared to the other two groups. However the shape of the chart in Figure 5.2 suggests that neither of the two older age groups would equal the interrupt performance of the 18-38 age group. When silence rate is considered the results indicate that despite a marginally significant ANOVA result, it cannot be confidently concluded that there is a significant correlation between age group and silence.

5.3.5. Verbal Ability

For the purposes of carrying out an analysis of variance with verbal ability as the categorical variable, the scores from the verbal section of the AH4 Group test of General Intelligence were converted into a binary distinction between High (H) and Low (L), the former being defined as the upper half of the distribution, the latter being defined as the lower half of the sample distribution. Figure 5.4 shows the mean number of interrupts across the four experimental trials for low and high verbal ability.

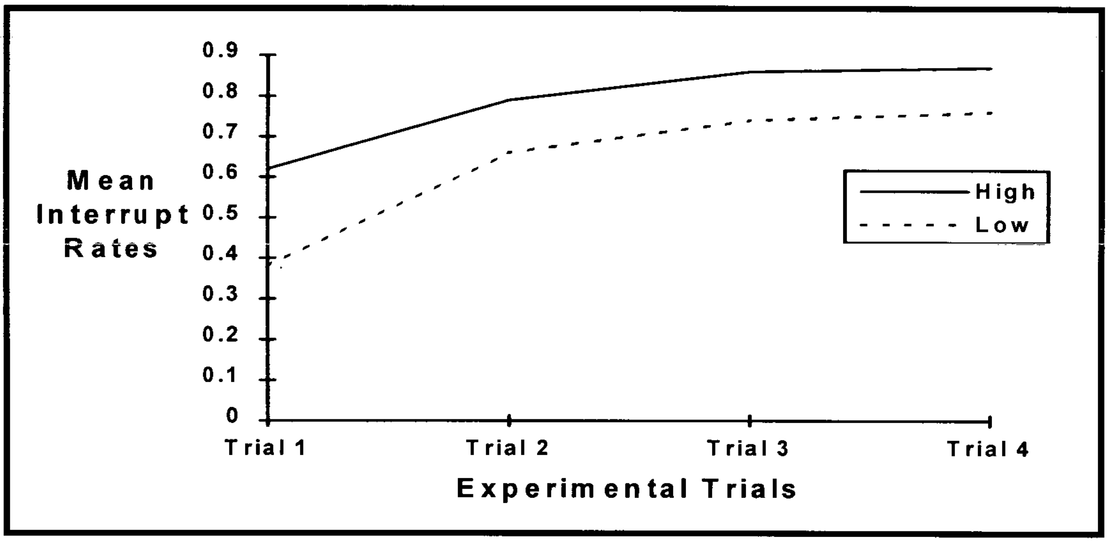


Figure 5.4. Mean Interrupt Rate for Verbal Ability

A significant difference was found for verbal ability across the four experimental trials, {F(1, 28) = 9.37, p < 0.01}. Related t-tests over trials 1 and 4 showed highly significant increases in use of interrupt by both low verbal

{t = -4.744, df = 15, p < 0.001} and high verbal ability {t = -8.225, df = 13, p < 0.001} subjects. A Tukey HSD test for trials 1 and 4 showed highly significant differences between low and high verbal ability subjects with respect to rate of interrupt {trial 1: $p < 0.01$; trial 4: $p < 0.01$}. In both cases, high verbal ability subjects interrupted more often than low verbal ability subjects.

Individual Difference # 11

Both low and high verbal ability users significantly increase their interrupt performance with keypad input after four uses of a service. However, even after four uses of a service, high verbal ability users were still significantly interrupting service prompts more often than low verbal ability subjects.

The number of times low and high verbal ability subjects were silent during the four experimental trials was also analysed. The data obtained for this measure are shown in Figure 5.5.

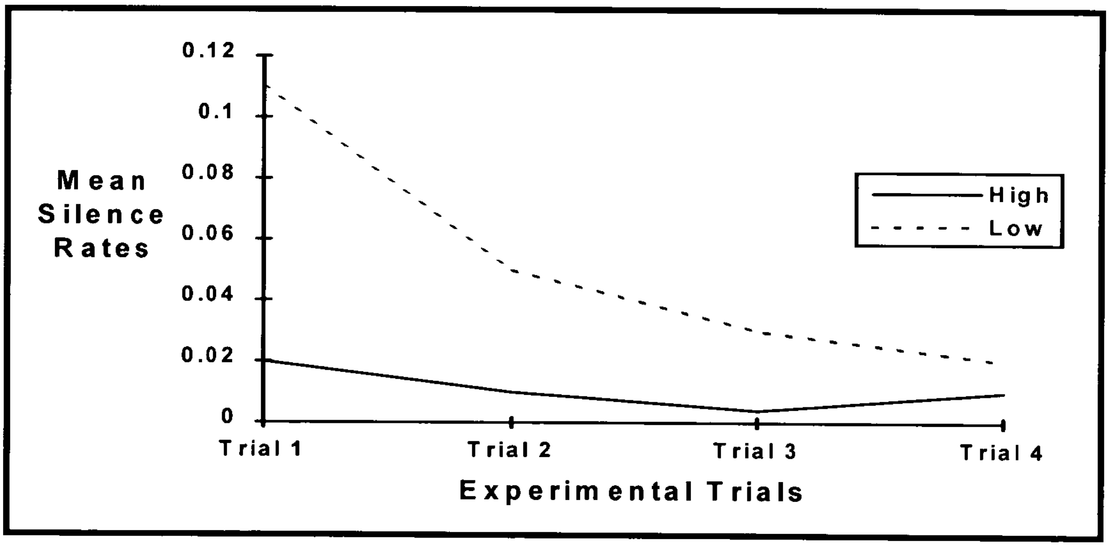


Figure 5.5. Mean Silence Rate for Verbal Ability

A significant difference was observed for verbal ability on the number of silences across the four experimental trials, {$F(2,27) = 5.42$, $p < 0.05$}. Related t-tests over trials 1 and 4 showed a significant decrease in silences by low verbal ability subjects {$t = 2.251$, $df = 13$, $p < 0.05$} but no significant difference was observed for the high verbal ability subjects. {$t = 0.611$, $df = 15$, $p = 0.551$}. A Tukey HSD test showed a significant difference between low and high verbal ability subjects for trial 1 ($p < 0.05$) in relation to silence rate with low verbal ability subjects having the higher silence rate. The difference between low and high verbal ability subjects' silence rates for trial 4 was not significant.

Individual Difference # 12

High verbal ability users have a significantly lower silence rate with keypad input than low verbal ability users after four uses of a service. Low verbal ability users significantly lowered their silence rate after four uses of a service.

When the Likert attitude data were analysed, no significant differences in attitude towards the automated music catalogue were found between low and high verbal ability subjects across the four experimental trials.

The results obtained from this analysis confirmed the results from the first experiment that there is a significant difference in the performance of low and high verbal ability subjects. Once again high verbal ability subjects perform better (i.e. higher interrupt rates, lower silence rates) than low verbal ability subjects, even if habituation to the automated music catalogue service is taken into account. It appears from the shape of the graph in Figure 5.5. that low verbal ability subjects would not reach the level of interrupt performance of high ability subjects for many trials, if at all.

5.3.6. Spatial Ability

As with the scores for verbal ability, the spatial ability sub-scale of the AH4 Group Test of General Intelligence was grouped into two categories - high and low - the former being defined as the upper half of the sample distribution scores for spatial ability and the latter being the lower half of the sample distribution. Figure 5.6 shows the mean number of interrupts across the four experimental trials for low and high spatial ability subjects.

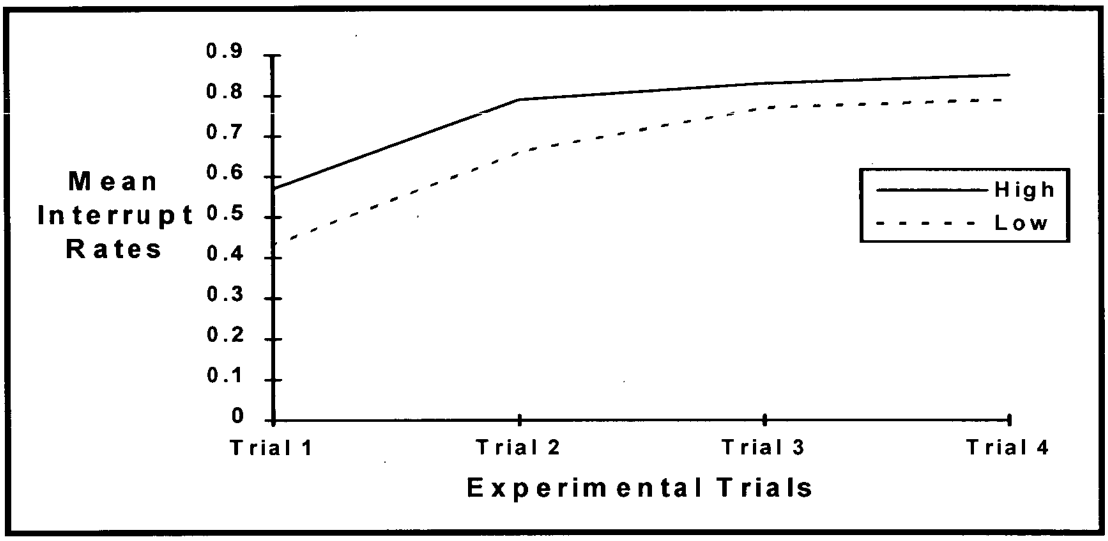


Figure 5.6 Mean Interrupt Rate for Spatial Ability

No significant differences were found between low spatial ability and high spatial ability subjects on interrupt rate across the four experimental trials. However, related t-tests over trials 1 and 4 showed highly significant increases in interrupt by both the low spatial ability subjects {t = -9.015, df = 12, p < 0.001} and the high spatial ability subjects {t = -4.842, df = 16, p < 0.001}. A Tukey HSD test showed no significant difference between high and low spatial ability subjects for either trial 1 or trial 4.

Individual Difference # 13

Both high spatial ability and low spatial ability users significantly increase their interrupt rate with keypad input after four uses of a service.

Differences between low spatial ability subjects and high spatial ability subjects on the occurrence of silences across the four experimental trials were also investigated. Figure 5.7 shows the mean number of silences across the four experimental trials for low and high spatial ability subjects.

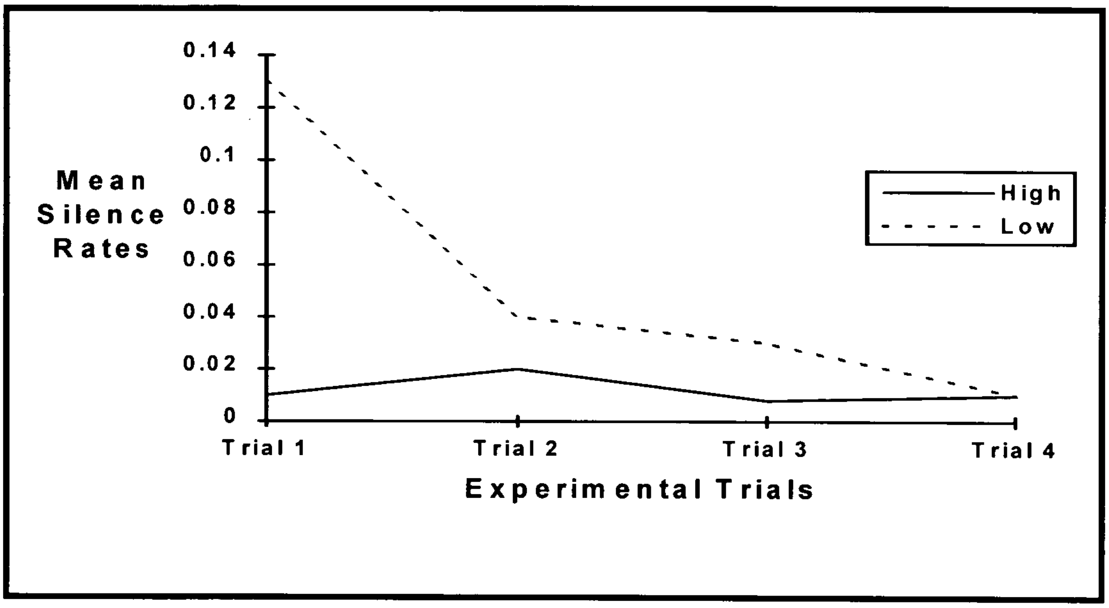


Figure 5.7 Mean Silence Rate for Spatial Ability

A significant difference in silence rate between low and high spatial ability subjects was found overall across the four experimental trials, {F(1, 28) = 4.60, p <0.05}. Related t-tests over trials 1 and 4 showed a significant decrease in silences by the low spatial ability subjects {t = 2.723, df = 12, p < 0.05} but no significant difference for the high spatial ability subjects. A Tukey HSD test showed a highly significant difference between low and high spatial ability subjects on the first trial (p < 0.01) but no significant differences between theses subjects on the 4 trial.

Individual Difference # 14

There is a significant difference in silence rate with keypad input between low and high spatial ability users. Low spatial ability users significantly improve upon their silence rates for keypad input after four uses of a service.

When differences in spatial ability were explored, no significant differences were found between low spatial ability and high spatial ability subjects with respect to attitudes towards the automated music catalogue.

The results for low and high spatial ability subjects follow a similar pattern to those obtained for verbal ability; with the most noticeable exception being that there was no significant difference overall between low and high spatial ability subjects in terms of interrupt rates. In addition, when these results are taken in conjunction with the results obtained for low and high spatial ability subject performance on the keypad version of the automated shopping catalogue, it would appear that spatial ability does not have as significant an effect on subject performance when using automated telephone services as verbal ability. This hypothesis is tested later in this chapter when a multiple regression analysis is carried out using these data.

5.3.7. Memory

For the purposes of carrying out an ANOVA with memory as the categorical variable, the scores from the subjects AVLT 1 to AVLT 5 were converted into a binary distinction between high and low, the former being defined as the upper half of the distribution and the latter being defined as the lower half of the sample distribution. Table 5.4 below shows the mean interrupt rate and table 5.5 shows the silence rates across the four experimental trials for low and high memory scores.

	AVLT 1		AVLT 2		AVLT 3		AVLT 4		AVLT 5	
Trial	Low	High	Low	High	Low	High	Low	High	Low	High
1	0.45	0.54	0.41	0.60	0.42	0.58	0.43	0.55	0.41	0.56
2	0.71	0.74	0.67	0.78	0.67	0.78	0.69	0.75	0.67	0.77
3	0.78	0.82	0.79	0.82	0.79	0.83	0.77	0.83	0.77	0.83
4	0.80	0.83	0.80	0.84	0.80	0.84	0.80	0.83	0.80	0.83

Table 5.4 Mean Interrupt Rate by Memory and Trial

A One-Way repeated measures ANOVA was performed on these data. No significant difference was found between low and high memory scores in relation to interrupt rate across the four experimental trials. The number of times subjects were silent after a prompt across the four experimental trials was also analysed in the same way. Once again, no significant differences were observed between low and high memory scores in relation to silence rate across the four experimental trials. The subjective measure analysis, based on the Likert questionnaire data, did not produce any significant difference in attitude across the four experimental trials between low and high memory scores.

	AVLT 1		AVLT 2		AVLT 3		AVLT 4		AVLT 5	
Trial	Low	High	Low	High	Low	High	Low	High	Low	High
1	0.08	0.06	0.08	0.05	0.07	0.06	0.10	0.05	0.10	0.04
2	0.03	0.03	0.04	0.03	0.03	0.03	0.05	0.02	0.04	0.02
3	0.03	0.01	0.03	0.01	0.03	0.01	0.04	0.01	0.04	0.01
4	0.02	0.01	0.03	0.01	0.03	0.01	0.02	0.01	0.02	0.01

Table 5.5 Mean Silence Rate by Memory and Trial

Overall there were no significant differences in performance in relation to memory scores on the AVLT. This seems quite surprising given the fact that subjects were required to make choices from menus of differing lengths throughout the interaction. One possible reason why there were no significant differences found in this experiment is that subjects had the information about the tracks they had to order in a sheet in front of them and any difficulty experienced when interacting with the service could be attributed to navigation issues. In addition, subjects had the opportunity to interrupt a system prompt therefore reducing the necessity of trying to hold all the information the system was giving them in their working memory before deciding the appropriate key to press.

5.3.8. Information Processing

For the purpose of carrying out an ANOVA with information processing as the categorical variable, subjects scores on both the 4 second and 2 second versions of the PASAT were converted into a binary distinction between high and low, the former being defined as the upper half of the distribution and the latter being defined as the lower half of the sample distribution. Table 5.6 shows the mean interrupt rate across the four experimental trials for low and high scores on the 2 second PASAT test.

PASAT Grouping	Trial 1	Trial 2	Trial 3	Trial 4
High	0.50	0.74	0.81	0.80
Low	0.52	0.72	0.80	0.85

Table 5.6 Mean Interrupt Rate for 2 Second PASAT Score and Trial

No significant difference was found between low scores and high scores on the 2 second PASAT test in interrupt rate across the four trials, F(1,28) = 0.05, p = 0.83. This analysis was then repeated using the scores obtained by subjects on the 4 second PASAT test. The data used in this analysis are shown in Table 5.7.

PASAT Grouping	Trial 1	Trial 2	Trial 3	Trial 4
High	0.47	0.74	0.83	0.83
Low	0.54	0.72	0.79	0.82

Table 5.7 Mean Interrupt Rate for 4 Second PASAT Score and Trial

No significant difference between low and high scores on the 4 second PASAT test in relation to the subjects' interrupt rate across the four experimental trials. The next data to be analysed were the silence rates across the four experimental trials, using the scores obtained from the PASAT tests.

No significant difference was found between low and high 2 second PASAT scores or between low and high 4 second PASAT scores in relation to silence rate across the four experimental trials.

When it came to the subjective measure analysis, once again there was no significant difference found between subjects who scored low and high on the PASAT test in their attitudes towards the automated catalogue service across the four experimental trials.

These results seems surprising as it might be expected that subjects who processed information quickly would have the information they needed to interrupt the system more and be silent less often than those subjects who processed information more slowly. On the other hand, it could be that the experimental task did not require a significant amount of cognitive effort on the part of the user. Supporting evidence for this can be seen in the results obtained from the AVLT analysis.

5.3.9. Gender Differences

The possibility of significant difference in performance due to gender was explored next, primarily once again with a One-Way repeated measures ANOVA. Table 5.8 shows the mean percentage interrupt rate across the four experimental trials for male and female subjects.

Gender	Trial 1	Trial 2	Trial 3	Trial 4
Male	0.55	0.71	0.81	0.83
Female	0.47	0.75	0.80	0.85

Table 5.8 Mean Interrupt Rate by Gender and Trial

The result of the analysis of variance indicated that there was no significant difference between males and females in their interrupt rates across the four experimental trials, {F(1,28) = 0.78, p = 0.78}. A similar type of analysis found no significant differences between males and females in terms of silence rate across the four experimental trials, {F(1,28) = 0.003, p = 0.96}.

The subjective measures analysis, again using a One-Way repeated measures ANOVA, indicated that there was no significant difference in male and female attitude towards the automated music catalogue across the four experimental trials. This reconfirmed the findings of the first experiment.

The results obtained from this experiment indicate that gender differences do not produce significantly different results in terms of performance and attitude towards the automated music catalogue. This finding is consistent with the gender differences results obtained in the first experiment.

5.3.10. Personality Differences

As with the scores for the other psychometric tests used in this experiment, the personality scores obtained from the NEO-PI-R were converted into a binary distinction between high and low, the former being defined as the upper half of the distribution and the latter being defined as the lower half of the sample distribution.

No significant differences in interrupt rate were observed between low and high groupings of subjects on the five personality traits across the four experimental trials. A similar set of findings was obtained when the silence rates for each of the five personality traits across the four experimental trials were considered.

When the Likert attitude data were analysed, no significant differences in attitude towards the automated music catalogue were observed between low and high personality trait scores across the four experimental trials. Although personality is regarded as the most stable of individual characteristics, the results indicate that personality differences do not lead to a significant variation in performance of a human-computer interaction task involving a system like the automated music catalogue. This finding also extends to attitude towards the service. The results are in line with most of the research into personality differences and performance on a human-computer interaction task.

5.4. Isolating Individual Differences

In step two of the three-stage methodology to investigate the effects of individual differences on human-computer interaction outlined in Chapter Three the text editing task was decomposed into three general components (finding, counting and generating) in order to isolate where the salient individual differences, identified by the "assay", have their greatest effect.

The task of using the automated music catalogue can also be decomposed into three components. As with the Egan and Gomez analysis, the first component is *finding*. In this case it is finding the appropriate music category, the appropriate artist, the appropriate album and the appropriate album track from the menu selections available. The next component is *generating*. In the case of the automated music catalogue it is generating the correct keystroke on the keypad from the menu choices available at the track level. At this stage of the interaction subjects may have to generate the correct keystrokes which allow them to listen to all three tracks available on an album before choosing to add one to their CD. The third component is *navigation*. After subjects have made a selection at track level they use menu 3 to navigate their way to another part of the music catalogue database. Whereas Egan and Gomez simulated their component processes by carrying out pencil and paper tasks, the actual data obtained from subjects' performance when using the automated music catalogue can be used to test this three component process hypothesis.

In order to explore the idea of component processes operating in subjects' interactions with the automated music catalogue, the menu structure of the automated music catalogue can be broken down as follows: finding: top level category, music category menu level, artist menu level and album menu level; generating: menu 1 and menu 2;navigation: menu 3.

The top level (TPL) is the first node in the hierarchically structured music database. The music category level (MCL) menu offers the subject five different categories of music from which to choose, e.g. Blues, Jazz, Classical, Rock & Pop, Folk. After making a selection at this level the subject goes down to the artist menu (ARTL). At this level the subject is presented with a list of artists from the music category they have chosen. After choosing an artist, subjects are then presented with an album menu (ALBL) for the artist of their choice and from a list of three they will make a selection. From the chosen album they will select from three tracks.

The next stage of the interaction involves the generating component. Menu 1 (M1) presents subjects with the options of playing the track they have just selected, adding it to their CD or making a different selection (choosing this option will send subjects to Menu 3). Menu 2 (M2) gives subjects the option of playing the track they have just heard once again, adding the track to their CD or making a different selection (choosing this option will also send subjects to Menu 3). Menu 1 and

Menu 2 represent the generating component part of the task because subjects are given the information which allows them generate commands to complete the task (i.e. select a track to put on a CD).

The third stage of interaction involves the navigation component. Menu 3 (M3) gives the subject the opportunity to list the three album tracks they can select from, list all the albums by the artist they have selected, or obtain a listing of all the artists of the music category they are currently in which allows them to move to another album by the same artist or move to a different artist within the same music category. Subjects also have the opportunity to move to other music categories. By choosing this last option, subjects will be sent back to the music category level and move on to the next task.

Having defined the component levels of the interaction, the next stage is to look for individual differences in performance at these specific levels. On the basis of the results already obtained, performance measures will be based on mean interrupt rates and mean silence rates. The individual differences assessed are age, verbal ability and spatial ability.

5.4.1. Component Differences in Interrupt Rates

Figure 5.8 shows the mean interrupt rates (averaged across the four experimental trials) at the specific menus for low and high verbal ability subjects. The labels on the x axis represent the following specific menus:

- TPL - top level menu
- MCL - music category menu
- ARTL - artist menu

- ALBL - album menu
- M1 - menu 1
- M2 - menu 2
- M3 - menu 3

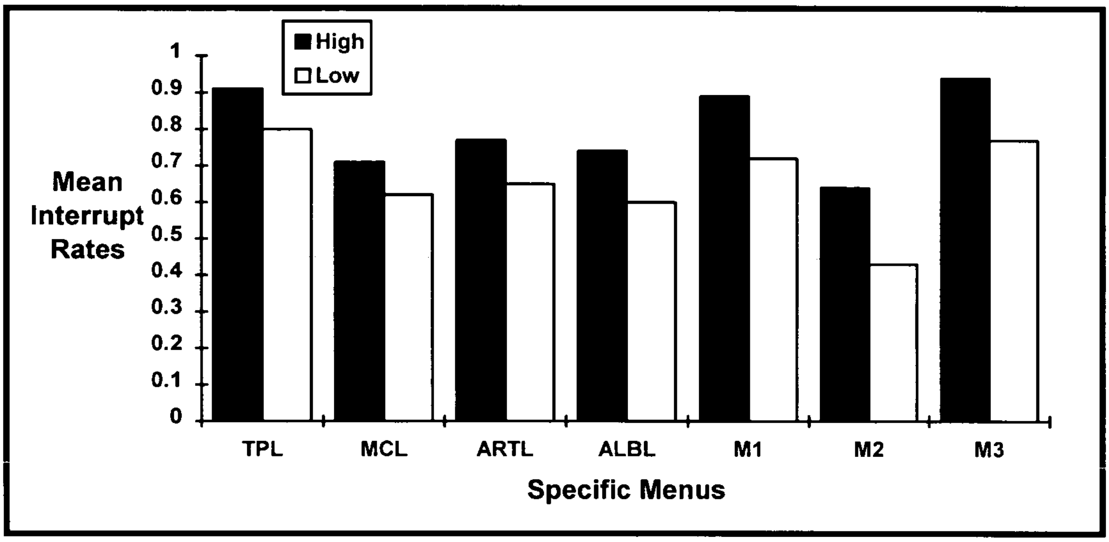


Figure 5.8 Mean Interrupt Rates for Specific Menus by Verbal Ability

A series of One-Way ANOVAs was performed on this data. A significant difference was observed between low and high verbal ability interrupt rates at the top level menu, {F(1,28) = 5.411, $p < 0.05$}, at the artist menu level, {F(1,28) = 7.175, $p < 0.05$}, at the album menu level, {F(1,28 = 9.156, $p < 0.001$}, at the Menu 1 level; {F(1,28) = 8.394, $p < 0.01$} and at Menu 3 level, {F(1,28) = 8.074, $p < 0.01$}.

Individual Difference # 15

The interrupt rate for high verbal ability users is significantly higher than low verbal ability users in the finding, generating and navigating component processes of an automated telephone service.

The next individual characteristic to be examined was spatial ability. Figure 5.9 shows the mean interrupt rates (averaged across the four experimental trials) at the specific menu levels for low and high spatial ability subjects.

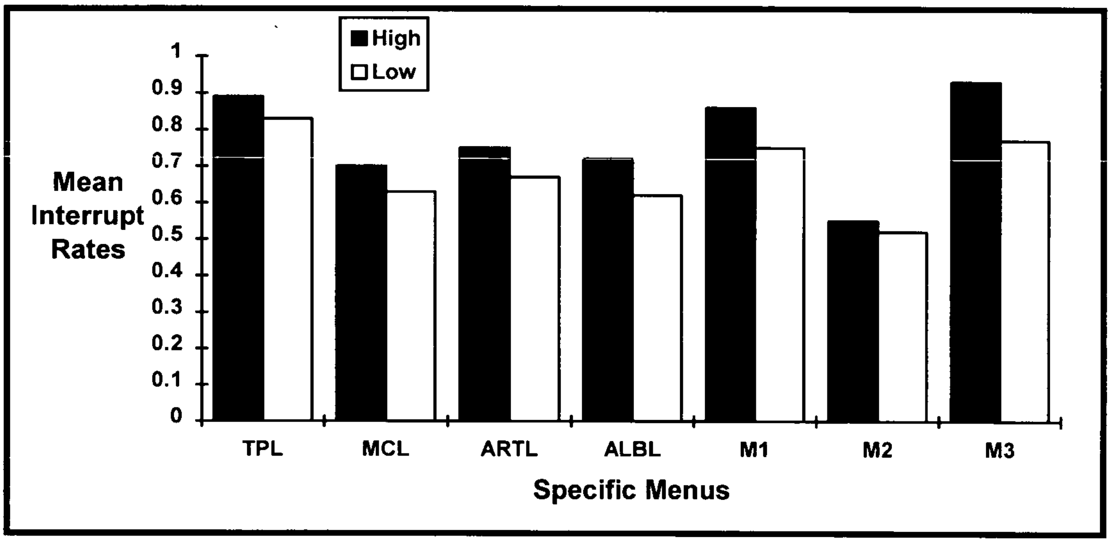


Figure 5.9 Mean Interrupt Rates for Specific Menus by Spatial Ability

A series of One-Way ANOVAs were performed on this data. A significant difference in interrupt rate between low spatial and high spatial ability subjects was observed only at Menu 3 level, {F(1,28) = 7.388, $p < 0.05$}.

Individual Difference # 16

High spatial ability users interrupt navigation prompts significantly more often than low spatial ability users.

Age differences in interrupt rate performance were examined next. Figure 5.10 shows the mean interrupt rate (across the four experimental trials) for the three groups.

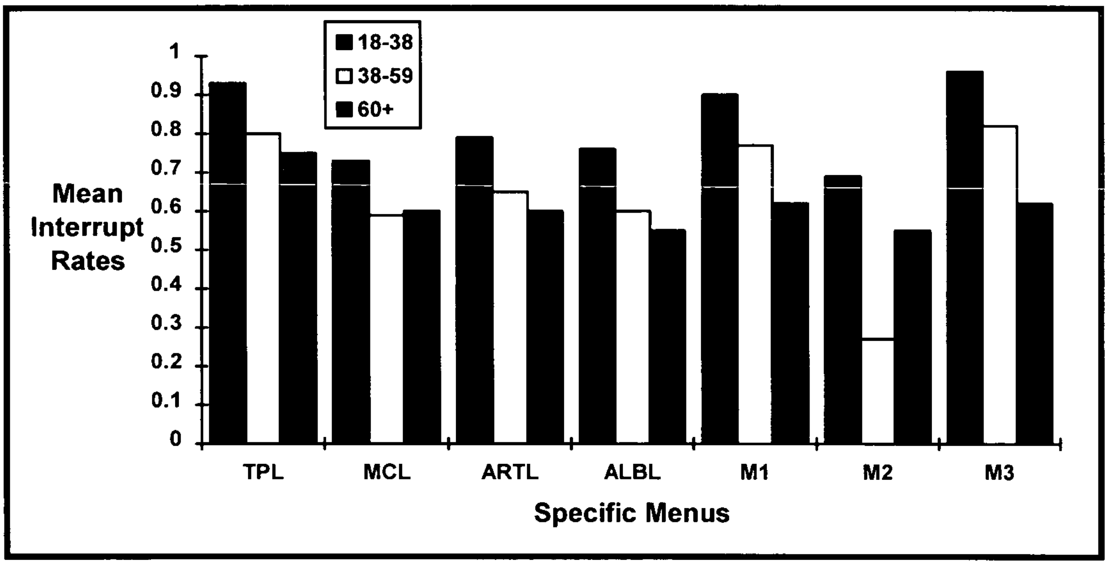


Figure 5.10 Mean Interrupt Rates for Specific Menus by Age

A series of One-Way ANOVA's were performed on these data. The results indicated that there was a difference in interrupt rates between the three age groups on every menu level: at the top level {F(2,27) = 5.70, p < 0.01}; music category level {F(2,27) = 3.75, p < 0.05}; artist level {F(2,27) = 7.66, p < 0.01}; album level {F(2,27) = 8.82, p < 0.001}; menu 1 {F(2,27) = 6.90, p < 0.01};

menu 2 {F(2,27) = 9.38, p < 0.01} and menu 3 {F(2,27) = 13.43, p < 0.001}. This suggests that age could have an effect on the generating, finding and navigation components in a task of this kind. This issue will be explored shortly.

Individual Difference # 17

There is a significant difference between the interrupt rates of all three age groups on all three component processes in an automated telephone service.

5.4.2. Component Differences in Silence Rates

Figure 5.11 shows the mean silence rate (across the four experimental trials) in specific menu levels for low and high verbal ability subjects.

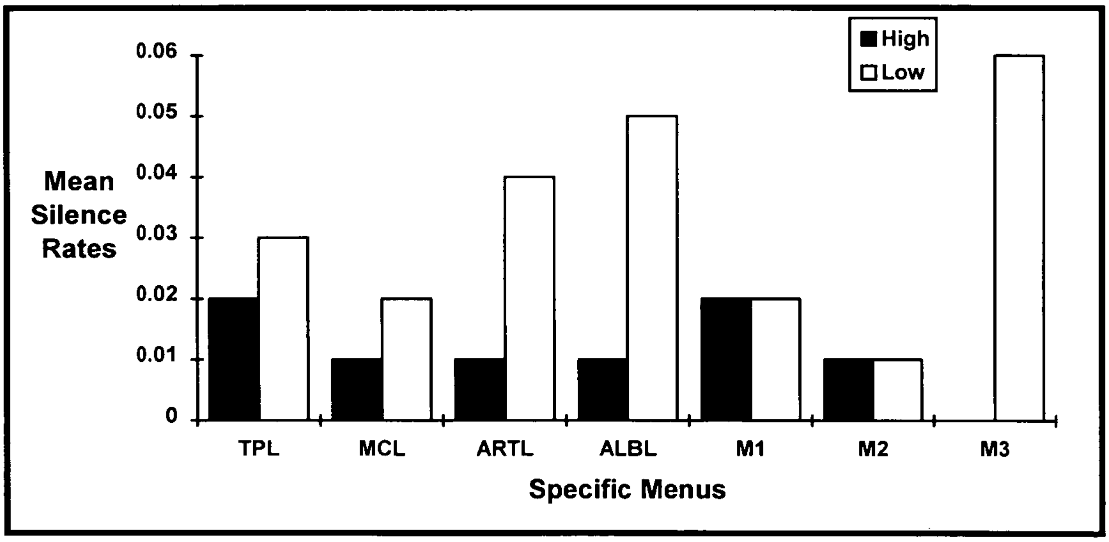


Figure 5.11 Mean Silence Rates for Specific Menus by Verbal Ability

A series of One-Way ANOVA's were performed on this data. A significant difference in silence rate was observed between low and high verbal ability subjects at the album level, {F(1,28) = 7.01, p < 0.05} and at the menu 3 lvel, {F(1,28) = 7.22, p < 0.05}.

Individual Difference # 18

There is a significant difference in silence rates on the finding and navigation processes in an automated telephone service for low and high verbal ability users.

The analysis was then repeated for low and high spatial ability subjects. Figure 5.12 shows the mean silence rates (across the four trials) at the specific menu levels for low and high spatial ability subjects.

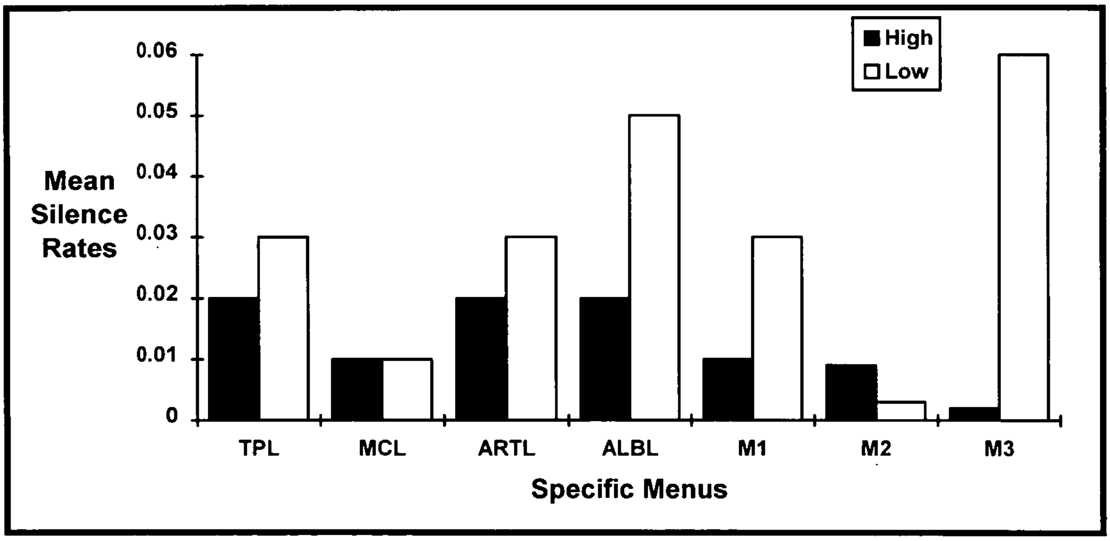


Figure 5.12 Mean Silence Rates for Specific Menus by Spatial Ability

A series of One-Way ANOVA's were performed on this data. A significant difference in silence rate between low and high spatial ability subjects was found at the menu 3 level, {F(1,28) = 8.58, p < 0.05}.

Individual Difference # 19

High spatial ability users have a significantly lower silence rate than low spatial ability users on the navigation process of an automated telephone service.

The final individual characteristic to be looked at was age. Figure 5.13 shows the mean silence rate (across the four experimental trials) at specific menu levels for age.

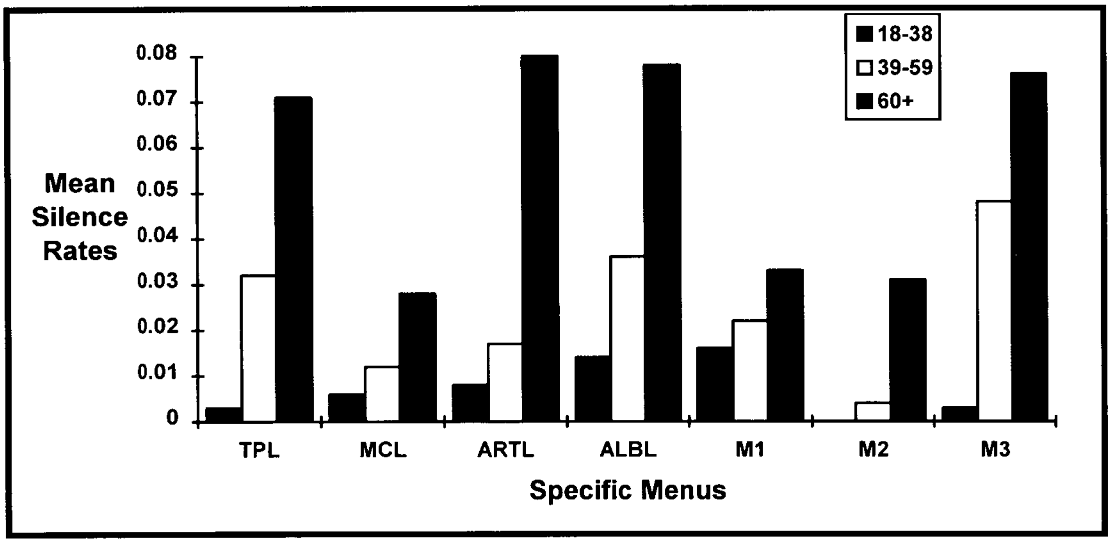


Figure 5.13 Mean Silence Rates for Specific Menus by Age

A series of One-Way ANOVA's were performed on these data. A significant difference was found for age at the top level menu {F(2,27) = 4.38, p < 0.05}; artist level {F(2,27) = 8.44, p < 0.01}; album level {F(2,27) = 5.65, p < 0.01} and at the Menu 2 level, {F(2,27) = 4.26, p < 0.05}.

Individual Difference # 20

There are significant differences in silence rates between all three age groups on the finding and generating processes of an automated telephone service.

The results obtained from the component analysis indicate that age and verbal ability are the two individual characteristics where the most significant differences occur. There are significant differences for both the finding and generating components when using silence rates and interrupt rates as the objective measures of performance and there were also significant differences observed between the age groups when the navigation component part of the task was looked at. Having identified the salient individual characteristics and indicated where these differences are most significant the next stage is to use the subjects' age and verbal ability scores in multiple regression equations to predict performance on each component task. Table 5.10 shows the standardised coefficients for age and verbal ability using silence mean interrupt rate as the criterion value for each component task.

Component	Age	Verbal Ability	R^2
Top Level	-0.46*	0.15	0.31**
Music Category	-0.26	0.18	0.11
Artist	-0.43*	0.18	0.29**
Album	-0.48*	0.18	0.35**
Menu 1	-0.43*	0.23	0.37**
Menu 2	-0.36	0.04	0.15
Menu 3	-0.56***	0.28	0.54***
Complete Task	-0.54**	0.17	0.40**

Table 5.10 Relationships among Component Tasks and Age and Verbal Ability Using Interrupt Rate as Criterion Measure

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

The multiple regression analysis was then repeated using silence rates for each of the component tasks as the criterion values. Table 5.11 shows the standardised coefficients for age and verbal ability and how they relate to component task performance.

Component	Age	Verbal Ability	R^2
Top Level	0.48*	0.09	0.20
Music Category	0.36	-0.04	0.15
Artist	0.49*	-0.02	0.25
Album	0.40*	-0.27	0.34**
Menu 1	0.05	-0.09	0.01
Menu 2	0.51*	0.23	0.20*
Menu 3	0.24	-0.39*	0.36**
Complete Task	0.32	-0.29	0.28*

Table 5.11 Relationships among Component Tasks and Age and Verbal Ability Using Silence Rate as the Criterion Measure

*$p < 0.05$, **$p < 0.01$

The results from the multiple regression analysis, taking into consideration both objective measures of performance, indicate that age had a significant effect on all three of the component processes - finding, generating and navigation - in the experimental task, whereas the effects of verbal ability appear to be confined to the navigation component part of the task only.

As in the Egan and Gomez approach, a tentative working hypothesis can be presented, based on these results. Figure 5.14 shows the working hypothesis relating age and verbal ability to the automated music catalogue component operations.

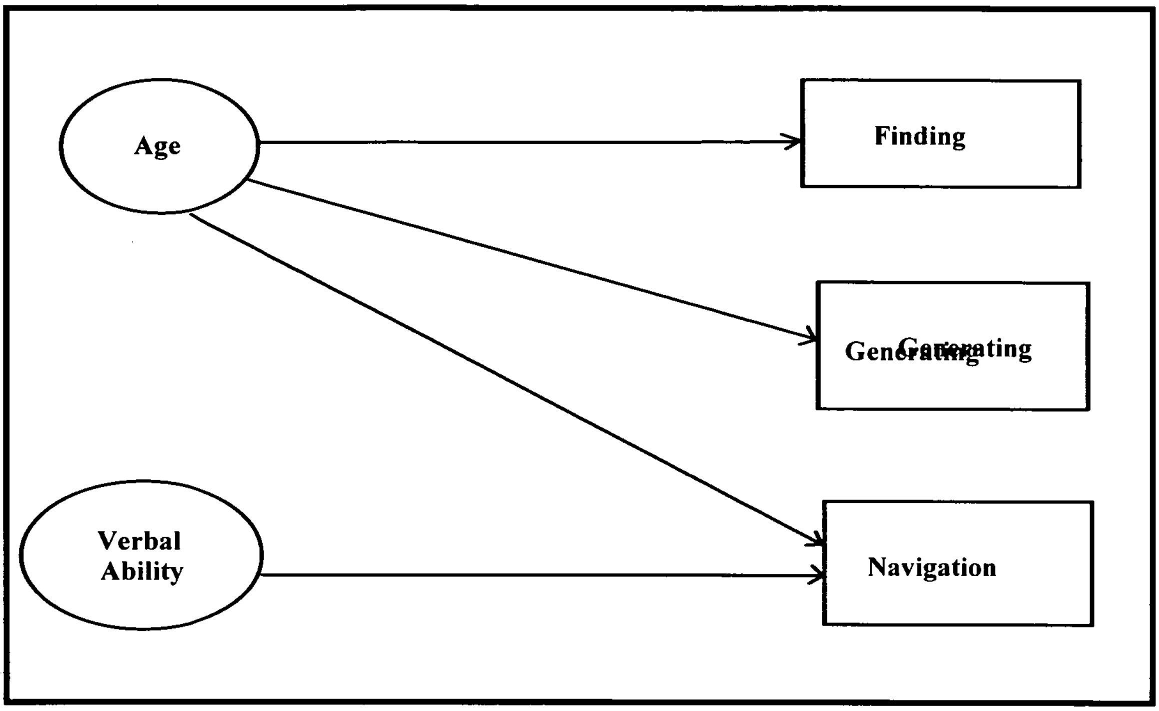


Figure 5.14 Working Hypothesis Relating Age and Verbal ability to Automated Music Catalogue Component Operations

Some corroborating evidence for this hypothesis comes from an examination of task completion rates across the four experimental trials. Considering age differences Table 5.12 shows that on only one occasion over the four experimental trials did any of the subjects in the 60+ age group successfully complete the task. The task completion frequencies for the youngest age group (18-38) indicate that this group of subjects did not experience the sorts of problem that the 60+ age group encountered.

Trial	Outcome 1		Outcome 2		Outcome 3		Outcome 4	
	18-38	60+	18-38	60+	18-38	60+	18-38	60+
1	10	0	5	1	0	1	1	3
2	11	0	4	1	0	4	1	0
3	11	0	5	3	0	1	0	1
4	13	1	2	4	0	0	1	0

Table 5.12 Frequency Table of Task Outcomes for Age Groups 18-38 & 60+

To recap, Outcome 2 refers to those subjects who did not follow the instructions but still ordered four tracks from the correct albums. Participants in this category may, for example, have chosen a track as required without listening to all the tracks available (as instructed). Outcome 3 refers to those subjects who did not follow the instructions and did not order tracks from the correct albums. Outcome 4 refers to those subjects who failed outright.

In terms of the hypothesis being put forward here, Outcome 2 errors reflect generating process problems on the part of the subject. After successfully finding the appropriate tracks these subjects cannot generate the correct key presses which will allow them to complete the task successfully (e.g. listen to all three tracks before making a selection). Outcome 3 and Outcome 4 indicate failures which combine the finding, generating and navigation processes. Here subjects either could not find the correct tracks or could not generate the correct key presses on the tracks that they did find. In addition, they had difficulty navigating their way through the automated music catalogue. The task completion rates indicate that older subjects appear to have more problems in all three component processes whereas the problems of younger subjects are confined more to the generating process.

5.5. Conclusions

The results obtained in this experiment confirm the 'assay' of the first experiment by once again showing significant differences in performance between the three age groups and between low and high verbal ability subjects. The results of the multiple regression analysis indicates that age and to a lesser extent verbal ability have a significant effect on subjects' performance when using an intelligent spoken language dialogue system like the automated music catalogue.

Following the approach of Egan and Gomez to isolating individual differences, the automated music catalogue task was broken down into three component processes - finding, generating and navigation. When individual differences were analysed, there were found to be significant differences between low and high verbal ability subjects and the three age groups for finding processes (e.g. selecting the appropriate artist from a particular music category), generating processes (e.g. using Menu 1 to select a particular track to play) and navigation processes (e.g. using Menu 3 to navigate to a different music category after selecting a track to put on the CD). The results obtained from the multiple regression analysis showed that age could be used to predict performance on finding, generating and navigation processes, whereas verbal ability was confined to predicting performance on navigation processes.

Another key aim was to explore the effect of learning on subjects' performance. There was a significant increase in the number of subjects who successfully completed the task between the first and the fourth trial, which suggests that improvement in subject's performance could be attributed to learning. Further evidence of the possible role played by learning can be seen in the significant increase in interrupt rate between the first and the fourth trial and the significant decrease in silence rate between the first and fourth trials. What is interesting from an individual differences perspective is that there were still significant differences in the interrupt rate performance between low and high verbal ability subjects and among the three age groups after the fourth experimental trial. A similar result was found when silence rates were analysed. The difference in performance becomes clearer when considering that there were nine subjects who never completed the experimental task at all over the four trials and the majority of these had low verbal ability and were in either the 38-59 or 60+ age groups.

The question which must now be addressed, taking the overall results into consideration, is how to accommodate for age, verbal ability and to a lesser extent spatial ability differences in order to bring the performance of low cognitive ability and older people closer to that of high verbal ability and younger subjects. The last stage of the Egan and Gomez methodology aims to accommodate the salient individual differences which have been identified in the first two steps through improvements like changes in the interface design and the instructions given to users before they use a system or service.

An important part of the research described in this thesis was a post-experience interview carried out with each subject. The results obtained from these semi-structured interviews indicated that although older and low verbal ability subjects felt as if their performance improved over the four uses of the automated music catalogue, they were still not completely satisfied with the way they had to interact with the service. As the ultimate aim of the Egan and Gomez methodology is to improve the usability of systems, subjects were asked what steps could be taken to improve the quality of their interaction with the automated music catalogue. The majority of these subjects stated that the various parts of the task (e.g. finding, generating and navigation) would be easier to complete if they could speak to the service. This was especially the case for the navigation component part of the task, which required users to make a decision about where they wanted to move in the database and select the appropriate key on the keypad to take them there. In addition, some of the older subjects expressed feelings of discomfort about using any form of new technology (e.g. personal computers, programming video cassette recorders) and stated a preference for the naturalness of speech over keypad entry for a task of this kind. Note that whilst results obtained in the first experiment suggest that older and low cognitive ability subjects express no difference in attitude towards keypad and connected word input, for a more complex task such as that involved when subjects are using the automated music catalogue, especially given the navigation component of this task, and users performance with menu 3,the desire for speech input is understandable..

Bearing this in mind, together with results obtained from the objective measures analysis, it was decided to try to accommodate for the needs of older and low verbal ability users by providing them a speech input version of the automated music catalogue.

Chapter 6: Accommodating Individual Differences

6.1 Introduction

The experiment reported in this chapter represents the third step (accommodating) of the methodology which has been used to assess the role of individual differences in spoken language dialogue systems. The results from the previous two experiments reported in this thesis have shown significant differences in performance between old and young subjects, and low and high verbal ability subjects when using automated telephone services like the automated music catalogue. In addition, the effects of these individual characteristics can be isolated to specific processes in the subjects' interaction with the system. The aim of this experiment therefore was to try to accommodate for these differences in order to provide a mode of interaction whereby the performance of older subjects would match that of the younger users.

This experiment required subjects to use two different versions of the automated music catalogue with the experimental task being similar to the one used in the experiment reported in the previous chapter of this thesis (i.e. subjects had to select 4 tracks of music to put on a 'personalised' CD). One version required subjects to use keypad input and the second required them to use speech input. The keypad version of the automated music catalogue service was the same as the one used in the previous experiment. The speech input version of the automated music catalogue had a simulated speech recognition level of 100% accuracy to allow a fair comparison between the two versions of the service. In the speech input version subjects would hear the same menu lists as those given by the keypad service, except at the end of a menu listing the service would prompt the user for speech input (e.g. "There are five categories of music available they are: Blues, Jazz, Classical, Rock & Pop and Folk. Please say which category of music you are interested in"). An options menu was also available for the user at every level of the user's interaction with the service but unlike the keypad version where users pressed the star key (*) in the speech input version subjects said "options". At the track level subjects could play or add a track to their CD by saying "play" or "add". As with the keypad version of the service, the speech input version of the automated music catalogue finished off the interaction with the user by playing back the artists and the tracks that had been ordered by the user and thanking them for using the service.

The success of an automated telephone service interface in accommodating the needs of a particular group of users is measured in the experimental work reported in this Chapter by examining two objective measures of performance (interrupt rate and silence rate) and one subjective measure of performance, namely the Likert attitude questionnaire. Taking interrupt rate first, it is generally regarded that a user will interrupt system prompts more often when they are confident that they know how to interact with the service in an effective and efficient way. Correspondingly, the less confident a user is with using a service the less likely they are to interrupt at the same rate as the confident user.

Silence can be regarded as a form of error when an individual is interacting with an automated telephone service like the automated music catalogue. If a user is unsure what to do at a particular stage of their interaction with the automated music catalogue they sometimes remain silent at the end of a prompt they have received from the system. Therefore silence may be a measure of a user failing to interact with the system in an appropriate way due to their failure to understand what they have to do at a specific point in their interaction with the automated telephone service.

The Likert attitude questionnaire used in this experiment was similar to the one used in the previous experiment. The wordings of two of the statements were changed when subjects completed the questionnaire related to the speech input version of the automated music catalogue: "I found it easy to make my responses using the telephone keypad" and "It was always clear which key I needed to press" were changed to "I found it easy to make my responses by speaking to the service" and "It was always clear what I needed to say" respectively.

As stated earlier, the main aim of this experiment was to investigate the possibility of speech input improving ('accommodating') the keypad performance of older and to a lesser extent low verbal ability subjects by bringing these groups of users' performance more in line with that of the younger and high verbal ability subjects. Also of interest was the attitude and performance of the younger and high verbal ability subjects with the speech input version of the automated music catalogue; in the first experiment (Chapter Four) their attitude towards the keypad version of the automated catalogue service was significantly better than their attitude towards the connected speech version of the service.

6.2 Methodology

6.2.1 Design

The experiment employed a repeated measures design, with subjects visiting the laboratory once to use two versions of the automated music catalogue: speech and keypad. Assignment of subjects to the experimental conditions was randomised to ensure that half of the subjects used the speech version of the service followed by keypad and half the subjects used the keypad version of the service first followed by the speech input version. For each version of the service, subjects were given a different list of music categories from which to make their selection. Each subject received the same list of items to order.

6.2.2. Subjects

In all, 40 subjects took part in this experiment, none of whom had previously used spoken language dialogue systems. Greater emphasis was placed on getting subjects in the 60+ age group to participate in the experiment, as the most significant results relating age to performance were found in this age grouping. The distribution of gender and age group is shown in Table 6.1.

	Ages 18-38	Ages 39-59	Ages 60+	Total
Male	6	2	11	19
Female	4	6	11	21
Total	10	8	22	40

Table 6.1 Age and Gender Breakdown of Subjects

6.2.3. Materials

The AH4 Test of General Intelligence (which provides a measure of verbal and spatial ability) was the only psychometric test given to the subjects in this experiment. Two versions of the Likert attitude questionnaire were used; one for each version of the automated music catalogue. In addition, an instruction sheet explaining how to use each version of the service was provided as well as a list of items to order with each version of the service.

6.2.4. Apparatus

The experimental set-up used in this experiment was similar to the two previous experiments. An IBM-compatible was set up in the room next door to the laboratory where the subject was located to run the Wizard of Oz software. A telephone was provided in the laboratory for subjects to use as part of the experiment.

6.2.5. Procedure

In this experiment the subjects attended the laboratory on one occasion. On arriving the subjects were given the AH4 Test of General Intelligence to complete. Following the completion of this test, subjects were given a list of instructions explaining how to use either the keypad or speech input version of the automated music catalogue and a list of items they were required to order. On completion of the task, each subject completed a Likert attitude questionnaire. After a short break, subjects were given a list of instructions explaining how to use the version of the service they had still to use and a list of items they were required to order. Once they had completed the task, they were given another Likert questionnaire to complete. The session finished with a semi-structured interview with subjects being asked specific questions like "if you had to use a version of the automated music catalogue again which version would you prefer to use and why would you prefer to use it?".

6.3. Results

The following measures of subjects' performance were made:

- successful task completion
- menu interrupt rate
- silence rate

In addition, the Likert attitude questionnaires provided subjective data on subjects' perceptions of the usability of the two versions of the automated music catalogue.

6.3.1. Task Completion Rates

The performance of subjects in terms of successful task completion between the two versions of the automated music catalogue (speech and keypad) is given in Table 6.2. Entries in columns under keypad and speech headings represent actual number of subjects.

Outcome	Keypad	Speech
1	4	6
2	17	19
3	13	8
4	6	7

Table 6.2 Subject Task Completion Performance

As in the experiment reported in Chapter five of this thesis, Outcome 1 refers to those subjects who followed the instructions exactly and completed all the tasks successfully; Outcome 2 refers to those subjects who did not follow the instructions but still ordered four tracks from the correct albums; Outcome 3 refers to those who did not follow the instructions and did not order four tracks from the correct albums and Outcome 4 refers to those subjects who failed outright.

The results outlined in Table 6.2 indicate that there was a slight improvement in overall task performance when subjects used the speech input version of the automated music catalogue but this improvement was not statistically significant. If the number of subjects who completed the task successfully are pooled with those subjects who did not follow the instructions exactly but completed the task, the results are similar here for both keypad and speech when compared to the first experimental trial of the second experiment. The overall failure rates for both the keypad and speech versions of the automated music catalogue were similar.

6.3.2. Interrupt Rate

Table 6.3 shows, for each version of the automated music catalogue, the mean percentage of user menu interrupts. Once again interrupt rate was calculated as the number of occasions a subject keyed in or spoke a response before the end of a system prompt divided by the overall responses to menu prompts and expressed as a percentage.

	Keypad	Speech
Mean Interrupt Rate	55%	46%

Table 6.3 Mean percentage interrupt rates for Keypad and Speech

The results in Table 6.3. indicate that subjects interrupted more, overall, with keypad as opposed to speech input. A related T-Test was performed on these data

which indicated that overall there was a significant difference in the interrupt rate between the two versions of the automated music catalogue,
{t = 2.87, df = 39, p < 0.001}, with subjects interrupting more when they used the keypad version of the service.

This finding was explored initially by comparing the interrupt rates between the two versions of the service in terms of order of exposure and then by comparing interrupt rates in terms of subject preference. Table 6.4 below shows the mean percentage interrupt rates for the two systems by order of exposure. Order of exposure (a) required subjects to use the keypad version of the automated music catalogue followed by the speech input version, and order (b) required the subjects to use speech input followed by keypad.

Order	Keypad	Speech
Keypad - Speech	54%	53%
Speech - Keypad	57%	40%

Table 6.4 Mean Percentage Interrupt Rate by Order of Exposure

A related T-test indicated that there was a significant difference in the interrupt rates between the two versions of the service for those subjects who used the speech input version followed by the keypad version of the service,
{t = 3.92, df = 19, p < 0.001}. No significant differences were found for those subjects who used keypad followed by speech. These results suggest that two independent effects could be present here: input mode and amount of previous experience. For example, the results shown in table 6.4 indicate that for subjects with no previous experience, those who use the keypad version of the service first interrupt more

often than those subjects who use the speech input version. Whereas subjects who used the keypad version of the service after the speech input version significantly increased their interrupt rates.

This analysis was then repeated for subject preference. Table 6.5 shows mean percentage interrupt rate by stated preference.

Stated Preference	Keypad	Speech
Keypad	59%	47%
Speech	52%	46%

Table 6.5 Mean Percentage Interrupt Rate by Preference

A Related T-test indicated that those subjects who indicated a preference for keypad (18 subjects out of 40) significantly interrupted this version of the service more than the speech input version, {$t = 2.27$, $df = 17$, $p < 0.05$}. Those subjects who indicated a preference for speech (22 subjects out of 40) actually interrupted more when using the keypad version of the service but not significantly. Therefore there does not appear to be a strong link between system preference and interrupt rate.

6.3.3. Silence Rate

The next data to be analysed was the silence rates for subjects when they used the two versions of the automated music catalogue. The percentage silence rate was calculated in a similar way to the percentage interrupt rate. Table 6.6 shows, for each version of the service, the mean percentage silence rates.

	Keypad	Speech
Mean Percentage Silence Rate	5%	1%

Table 6.6 Mean Percentage Silence Rate for Keypad and Speech

The data indicate that for both keypad and speech input, the level of silence rates is low. This result, for both keypad and speech input versions of the automated music catalogue, is similar to the result obtained for keypad in the first experimental trial of the music catalogue experiment reported in chapter five of this thesis. A related T-test indicated that, overall, subjects were silent significantly less when they used the speech input version of the automated music catalogue, {$t = 4.29$, $df = 39$, $p < 0.001$}. Therefore, it would appear that the speech input version of the automated music catalogue resulted in subjects making significantly fewer silence errors than the keypad version of the service.

As with the overall interrupt data, the silence data were also analysed in terms of order of exposure and preference. Table 6.7 shows the mean percentage silence rates for the two versions of the automated music catalogue in relation to order of exposure.

Order	Keypad	Speech
Keypad - Speech	5.9%	4.7%
Speech - Keypad	5.2%	4.6%

Table 6.7 Mean Percentage Silence Rate by Order of Exposure

A One-Way ANOVA (unrelated) showed that there was no significant difference in silence for keypad between the two orders of exposure, {F (1,38) = 0.759, p = 0.389}. A similar result was obtained for speech input, {F (1, 38) = 1.900, p = 0.176}. However, when a within subjects analysis was conducted using Related T-tests it was found that those subjects who used the keypad service followed by the speech input version were silent significantly more often when they used the keypad version, {t = 3.62, df = 19, p < 0.01}. A similar finding was also observed for those subjects who used the speech input version followed by the keypad version of the automated music catalogue, {t = 2.44, df = 19, p < 0.05}.

The silence rate data were then analysed in terms of subject system preference. Table 6.8 shows the mean percentage silence rate by subject preference.

Stated Preference	Keypad	Speech
Keypad	4%	1.2%
Speech	6%	0.8%

Table 6.8 Mean Percentage Silence Rate by Preference

Once again, the results of subject preference indicates that subjects' preference for a system does not always reflect their effectiveness and efficiency when using their preferred version of the automated catalogue service. A within subject analysis, using a Related T-test, indicated that subjects who preferred keypad to speech were significantly more silent when they used the keypad version of the automated music catalogue, {t = 2.39, df = 17, p < 0.05}. Those subjects who preferred speech input to keypad were also significantly more silent when they used the keypad system, {t = 3.59, df = 21, p < 0.01}. Once again, the results point towards speech being the more effective interface than keypad in terms of the number of silence errors subjects make when interacting with the system.

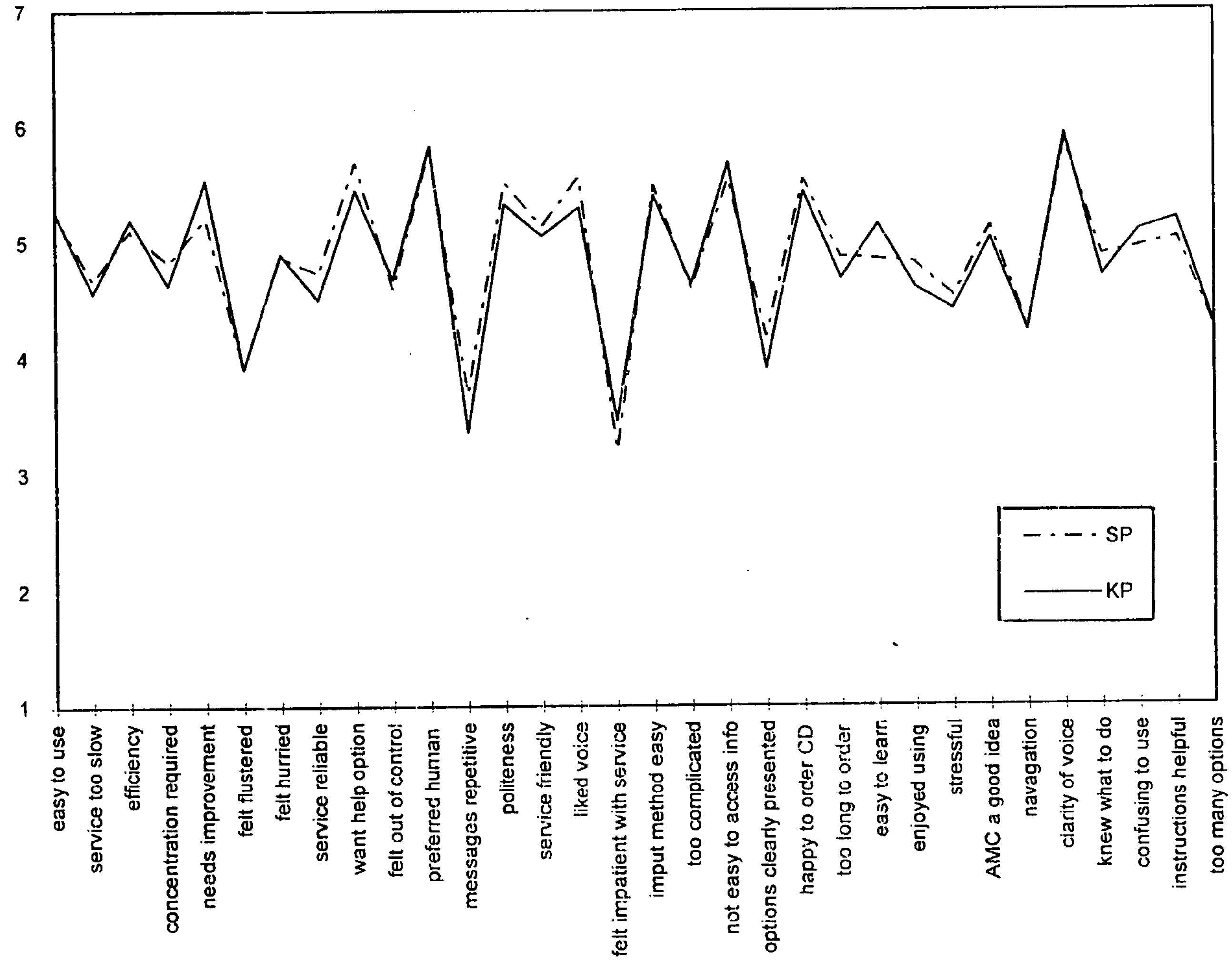
6.3.4. Summary of Interrupt and Silence Rate Results

The results obtained from the overall interrupt rates indicate that subjects interrupted more when they used the keypad version of the automated music catalogue rather than the speech input version. This was the case regardless of order of exposure to the system and also in contradiction, for some subjects, to their preferred mode of input. It appears, for interrupt rate, that mode of input and habituation have a role to play. However when silence rates are looked at a slightly different picture emerges. Habituation does not appear to influence subjects performance on either version of the automated music catalogue, whereas mode of input seems to be more influential. For example, in contrast to the interrupt rate results, subjects who use speech input first are silent significantly more often when they use the keypad version of the automated music catalogue. These findings are explored in more detail when the question of accommodating individual differences is addressed later on in this chapter of the thesis.

6.3.5 Attitude Scores

Figure 6.1 shows the mean attitude scores for the keypad and speech versions of the automated music catalogue.

Figure 6.1
Mean Subject Attitude Scores Across Input Modes



The mean attitude scores of subjects towards the keypad and speech input versions of the automated music catalogue are almost identical. A Related T-Test was performed on the data and confirmed that there was no significant difference between subjects' attitudes towards the keypad and speech versions of the automated music catalogue. This result can be contrasted with the finding obtained in the first experiment conducted as part of this thesis (reported in Chapter Three) which indicated that subjects prefer keypad to speech input - although this difference can be explained by the simple nature of the task the subjects were required to carry out in the first experiment, as opposed to the more complex nature of the task reported in this chapter.

Subjects' attitudes towards the two versions of the service were analysed in terms of order of exposure and subject preference. Table 6.9 shows the mean attitude score for speech input and keypad entry for order of exposure.

Order	Keypad	Speech
Keypad - Speech	4.88	4.99
Speech - Keypad	4.82	4.76

Table 6.9 Mean Attitude Score by Order of Exposure

The results for attitude are somewhat conflicting. Subjects who experienced speech input first had a more positive attitude towards the keypad version of the automated music catalogue, whereas those subjects who used keypad first had a more positive attitude towards the speech input version of the service. When a statistical analysis was carried out on these data, no significant difference was found between subjects' attitude towards the keypad and speech input versions of the automated music catalogue.

A similar analysis was then carried out for subject preference. Table 6.10 shows the mean attitude score for speech input and keypad entry in terms of subject preference.

Stated Preference	Keypad	Speech
Keypad	4.91	4.63
Kpeech	4.81	5.08

Table 6.10 Mean Attitude Scores by Preference

A Related T-test analysis was performed on this data. The results indicated that those subjects who stated a preference for keypad had a significantly more positive attitude towards it than speech input, {$t = 2.40$, $df = 17$, $p < 0.05$}. A similar result was observed for those subjects who indicated a preference for speech; they significantly preferred it to keypad entry in terms of attitude, {$t = 2.56$, $df = 21$, $p < 0.05$}. This finding was explored in more detail by examining where these differences in attitude occurred. As a result of this analysis it was found that those subjects who indicated a preference for speech felt they had to concentrate more when they used the keypad system {$t = 3.08$, $df = 21$, $p < 0.01$} and found that they got more flustered when they used the keypad system {$t = 2.51$, $df = 21$, $p < 0.02$}. On the other hand those subjects who expressed a preference for keypad entry felt it was easier to make responses using the telephone keypad {$t = 3.856$, $df = 17$, $p < 0.001$} and found the keypad version of the automated music catalogue less confusing to use than the speech input version {t = 3.06, df = 17, p < 0.01}.

6.4. Individual Differences

As stated, the main hypothesis of this experiment was that older and low verbal ability subjects would improve upon their keypad performance of the automated music catalogue by using a speech version of the system. The analysis which follows will also seek to confirm the model of subjects' interaction with an automated music catalogue which was put forward in the previous chapter (Chapter Five) of this thesis.

6.4.1. Verbal Ability Differences

Table 6.11 shows the mean interrupt rate for low and high verbal ability subjects across the two versions of the automated music catalogue.

Verbal Ability	Keypad	Speech
High	0.59	0.56
Low	0.51	0.36

Table 6.11 Mean Interrupt Rate for Verbal Ability

Table 6.11 shows that high verbal ability subjects interrupt more than low verbal ability subjects on both the keypad and speech input versions of the automated music catalogue. These differences were not significant. A One-Way ANOVA indicated that there was no significant difference between the interrupt rate of high and low verbal ability subjects when using either the keypad version of the automated music catalogue (p = 0.35) or the speech input version (p = 0.07).

However a Related T-test analysis revealed that low verbal ability subjects significantly interrupted the keypad version of the service more than the speech input version of the automated music catalogue, {t = 3.14, df = 18, p < 0.01}. No significant differences in interrupt rate for the two versions of the service was observed for high verbal ability subjects.

Individual Difference # 21

Low verbal ability users interrupt service prompts significantly more often when using keypad input than when using speech input.

This result appears to go against the experimental hypothesis as regards speech input improving low verbal ability subjects' performance, however when the overall silence rates are considered for low and high verbal ability subjects a slightly different picture begins to emerge. Table 6.12 shows the mean silence rate for the two versions of the automated music catalogue in terms of verbal ability.

Verbal Ability	Keypad	Speech
High	0.04	0.004
Low	0.06	0.02

Table 6.12 Mean Silence Rate by Verbal Ability

A One-Way ANOVA indicated that there was no overall difference in silence rate between low and high verbal ability subjects for keypad (p = 0.23) or speech input (p = 0.16) versions of the automated music catalogue. However when silence rate was compared within subject groupings by a Related T-Test it was found that when using the speech version of the automated music catalogue low verbal ability subjects were significantly less silent compared to when they used the keypad version of the service, {t = 3.02, df = 18, p < 0.01}. A similar result was also observed for the high verbal ability subjects, {t = 3.02, df = 20, p < 0.01}. It would appear that speech input accommodates both high and low verbal ability subjects in terms of silence rate but not interrupt rate. This result is explored in more detail when the effects of verbal ability are considered in the next section on performance at the component process level of interaction.

Individual Difference # 22

Low and high verbal ability users fail to provide a response to service prompts on fewer occasions when using speech input in comparison to keypad input.

The attitudes of low and high verbal ability subjects towards the two versions of the automated music catalogue were next to be analysed. Table 6.13 shows the mean attitude score for low and high verbal ability subjects towards keypad and speech.

Verbal Ability	Keypad	Speech
low	4.75	4.64
high	4.94	5.09

Table 6.13 Mean Attitude Score for Verbal Ability

The results in Table 6.13 show that the attitudes of low verbal ability and high verbal ability subjects are similar for both the keypad and speech input versions of the automated music catalogue. A One-Way ANOVA was performed on the attitude data. No significant difference in attitude was observed between low and high verbal ability subjects towards keypad entry (p = 0.55) or speech input (p = 0.18). In addition, the results of the within subjects analysis indicated that low and high verbal ability subjects had no significant difference in attitude towards the two versions of the automated music catalogue.

6.4.2. Age Differences

Table 6.14 shows the mean overall interrupt rates for age groups across the two versions of the automated music catalogue.

Age Group	Keypad	Speech
18-59	0.73	0.69
60+	0.40	0.28

Table 6.14 Mean Interrupt Rate by Age group

For the purposes of carrying out statistical analysis on the data for age, the results of the two age groups 18-38 and 39-59 were pooled. A One-Way ANOVA indicated that there was a significant difference in interrupt rates overall between the two age groups, {F(1,38) = 22.64, p < 0.001}, for keypad entry. A similar finding was observed when the interrupt rates of the two age groups for speech input were looked at, {F(1,38) = 12.40, p < 0.001}. A Related T-test showed that the younger subjects interrupted more when using the keypad version of the automated music catalogue than the speech version, {t = 2.97, df = 17, p < 0.01}. It was also shown that older people interrupted more when they used the keypad version of the service, {t = 4.27, df = 39, p < 0.001}.

Individual Difference # 23

The interrupt rate for younger users interrupt rate is significantly higher than for olde users for both keypad and speech input modes.

Again it appears as if speech is not accommodating the group of the older users. However, when the overall silence rates for age are looked at the reverse finding which was observed for verbal ability subjects is also present.

Table 6.15 shows the mean silence rate for both age groupings for the two versions of the automated music catalogue.

Age Group	Keypad	Speech
18-59	0.03	0.01
60+	0.07	0.01

Table 6.15 Mean Silence Rate by Age

A One-Way ANOVA was performed on this data. A significant difference in silence rate between the two groups when using the keypad version of the automated music catalogue was observed, {$F(1,38) = 4.44$, $p < 0.05$}. However when the two groups were compared for silences on the speech input version of the service, no significant differences in silence rate was observed, {$F(1,38) = 0.15$, $p = 0.70$}. This gives the first indication that speech has in fact accommodated older subjects in terms of matching their performance, as measured by silence rate, to other users. A related T-test indicated that subjects in the younger age group were significantly less silent when using the speech version of the service, {t = 3.06, df = 17, p<0. 01}, as were the older subjects, {t = 3.65, df = 21, p < 0.01}.

Individual Difference # 24

Younger users fail to provide a response to service prompts on significantly fewer occasions than older users when using keypad input.

These apparently contradictory findings in performance measures are explored in more detail in the next section of results, which looks for age differences at the component process level for the two versions of the automated music catalogue.

Attitude towards the two versions of the system was then considered. Table 6.16 shows the mean attitude scores for the two age groups for keypad and speech.

Age Group	Keypad	Speech
18-59	4.86	4.89
60+	4.85	4.86

Table 6.16 Mean Attitude Score by Age

The analysis of the attitude data for the two age groupings indicated that there were no significant differences between the two groups towards either the keypad or speech input of the automated music catalogue. Also no significant within subject differences were observed. This finding is similar to the one obtained in the experiment reported in Chapter 5 of this thesis where attitude towards a keypad version of the automated music catalogue was assessed.

6.4.3. Component Process Differences

6.3.4.1 Verbal Ability

Figure 6.2 shows the mean interrupt rate at the specific menu levels for high and low verbal ability subjects when they are using both the keypad and speech versions of the automated music catalogue.

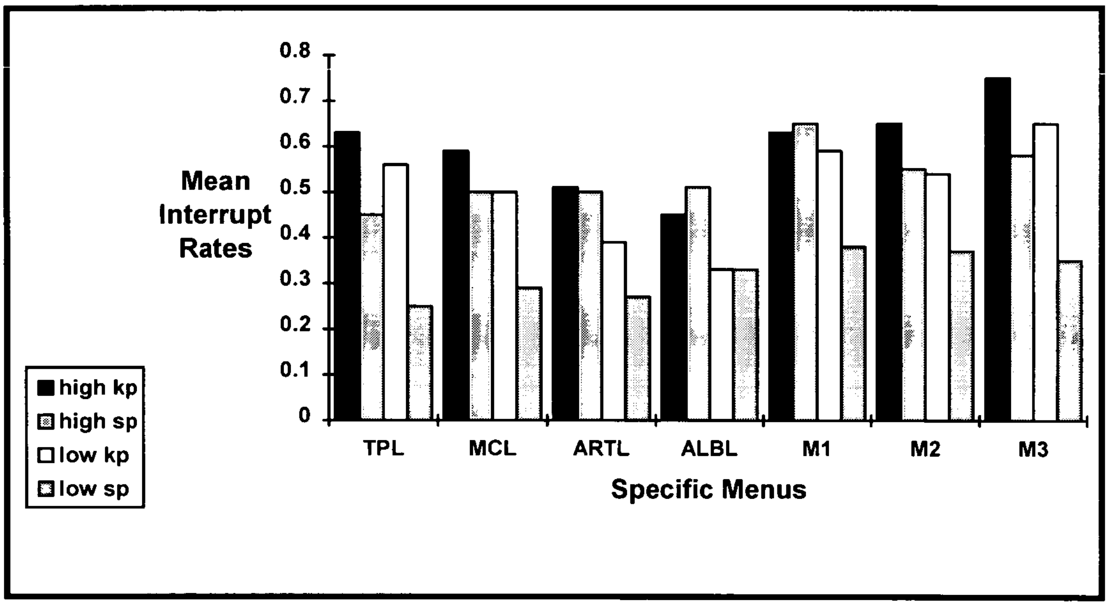


Figure 6.2 Mean Specific Menu Interrupt Rates by Verbal Ability

The key to the graph shown in figure 6.2 is as follows:

- **TPL** - top level menu
- **MCL** - music category menu
- **ARTL** - artist menu
- **ALBL** - album menu
- **M1** - menu 1
- **M2** - menu 2
- **M3** - menu 3
- **high kp** - high verbal ability subjects' mean interrupt scores for keypad
- **high sp** - high verbal ability subjects' mean interrupt scores for speech
- **low kp** - low verbal ability subjects' mean interrupt scores for keypad
- **low sp** - low verbal ability subjects' mean interrupt scores for speech

A One-Way ANOVA indicated that there were no significant differences in interrupt rate between low and high verbal ability subjects at the top level menu (TPL) for both keypad and speech versions of the automated music catalogue. There were significant within subject differences. A Related T-test indicated that high verbal ability subjects interrupted more at the top level of the automated music catalogue when they used keypad entry, {t = 2.88, df = 20, $p < 0.01$}. This was also the case for low verbal ability subjects, {t = 4.56, df = 18, $p < 0.001$}. Low verbal ability subjects also significantly interrupted more at the album level (ALBL) of the automated music catalogue when using keypad, {t = 2.92, df = 18, $p < 0.01$}. A significant difference in interrupt rate between low and high verbal ability subjects was also observed at the menu 1 level (M1) of the automated music catalogue when subjects were using the speech input version, {F(1,38) = 5.26, $p < 0.05$}. This was also the case at the menu 3 level (M3) for low and high verbal ability subjects with the speech version of the service, {F(1,38) = 4.86, $p < 0.05$}. Within subjects differences were also observed here with high verbal ability subjects interrupting more at this level when using keypad entry, {t = 2.55, df = 20, $p < 0.05$}. This was also true for low verbal ability subjects, {t = 3.52, df = 18, $p < 0.01$}.

Individual Difference # 25

Low and high verbal ability users interrupt menus in an automated telephone service significantly more often when they use keypad input mode when compared to speech input.

The overall results for interrupt rates indicated that both low and high verbal ability subjects interrupted more often when they used the keypad version of the automated music catalogue. The specific menu analysis has now shown that this increased interrupt rate for keypad is present in all three component processes. In addition, it appears to be the case that low verbal ability subjects will not interrupt at the same rate as high verbal ability subjects on either the keypad or speech input version of the automated catalogue service.

The silence rate for low and high verbal ability subjects at the component process level was analysed. Figure 6.3 shows the mean silence rate for the specific menu levels by verbal ability for both versions of the automated music catalogue.

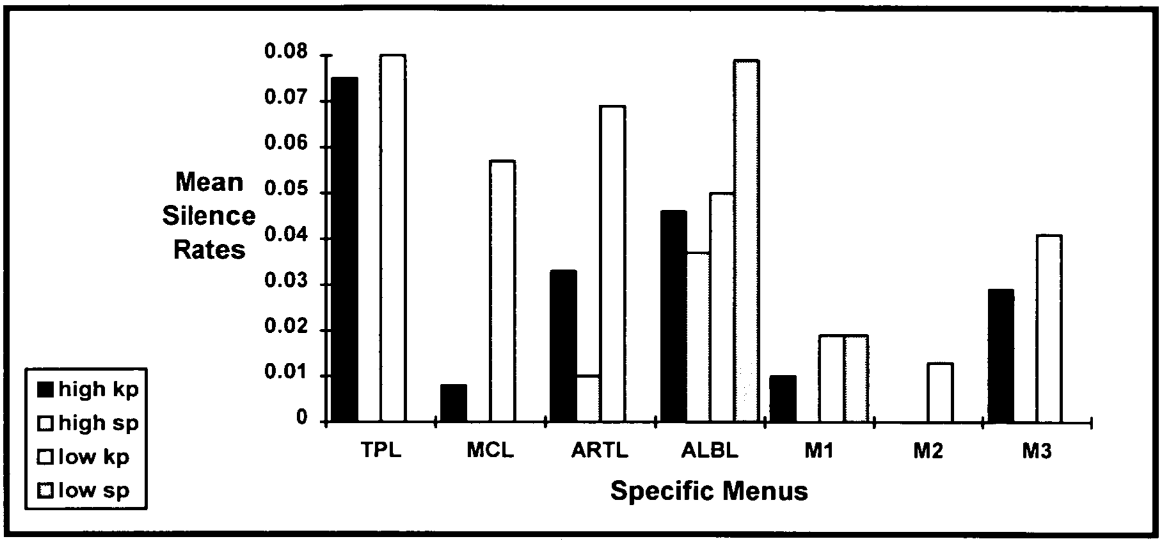


Figure 6.3 Mean Silence Rate for Specific Menus by Verbal Ability

Figure 6.3 shows that there was no great difference in silence rates at the menu levels between low and high verbal ability subjects for either the keypad or speech input versions of the service. The one significant exception was at the music category level (MCL). Here a One-Way ANOVA indicated that there was a significant difference in the silence rate between low and high verbal ability subjects at the music category level of the automated music catalogue when they were using keypad entry, {F(1,38) = 5. 02, p < 0.05}. A related T-Test indicated that low verbal ability subjects were significantly less silent at the music category level when they used the speech input version of the service, {t = 2.79, df = 18, p = 0.012}.

The difference in silence rates between low and high verbal ability subjects appears to be confined to the finding component process level which requires the subject to find the music category they are interested in, select the appropriate artist from that category, choose the correct album of the artist and select a track. In addition, both low and high verbal ability subjects make fewer silence errors when they use the speech input version of the automated music catalogue.

6.3.4.2. Age

For the age analysis at the component process level, the results of the age groups 18-38 and 39-59 were pooled. Figure 6.4 shows the mean interrupt rate for the specific menu levels for the two age groups.

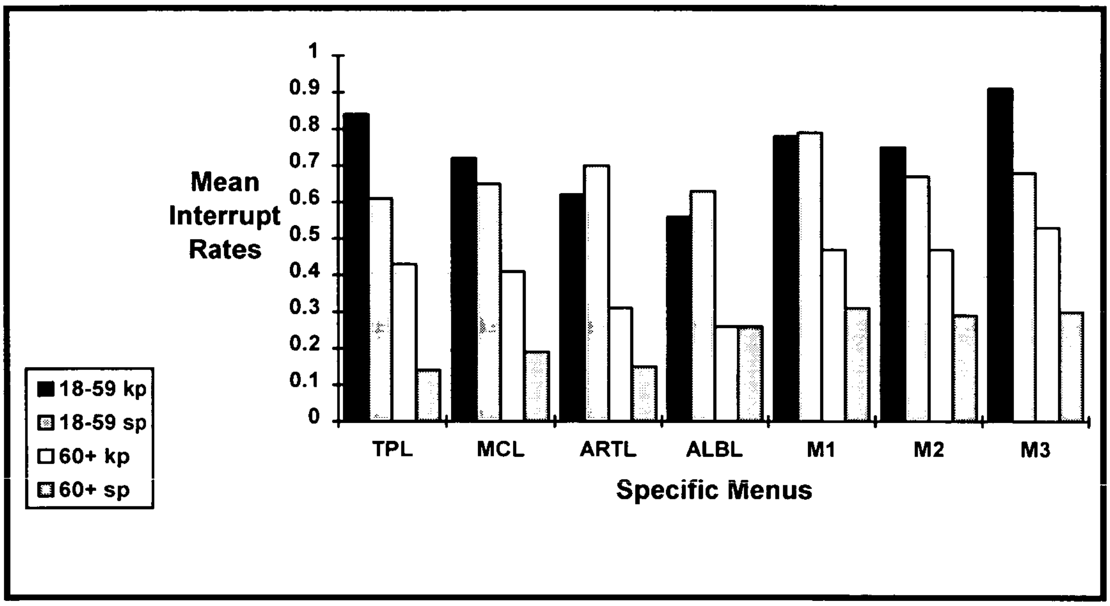


Figure 6.4 Mean Specific Menu Interrupt Rates by Age

The results of the analysis of differences in interrupt rates at the specific menu for the two age groups are summarised in Table 6.17.

Specific Menu Level	Age v Keypad	Age v Speech
top	p < 0.001	p < 0.001
music category	p < 0.05	p < 0.001
artist	p < 0.001	p < 0.001
album	p < 0.001	p < 0.001
menu 1	p < 0.01	p < 0.001
menu 2	p < 0.05	p < 0.001
menu 3	p < 0.001	p < 0.001

Table 6.17 Significance Levels for Age Differences on Interrupt Rate

As can be seen from table 6.17, there are significant differences between the two age groupings for each level of the automated music catalogue.

Individual Difference # 26

Younger users have a significantly higher interrupt rate than older userss on the finding, generating and navigation processes of an automated telephone service for both speech input and keypad input.

The within subject analysis for the specific menu levels revealed a similar set of results with both age groups interrupting significantly more often when using the keypad version of the automated music catalogue. The results of the within subject analysis is given in table 6.18

	18-59	**60+**
Specific Menu Level	**Keypad v Speech**	**Keypad v Speech**
top	$p < 0.01$	$p < 0.001$
music category	$p = 0.423$	$p < 0.01$
artist	$p = 0.356$	$p < 0.05$
album	$p = 0.360$	$p = 0.988$
menu 1	$p = 0.964$	$p < 0.05$
menu 2	$p = 0.469$	$p < 0.01$
menu 3	$p < 0.001$	$p < 0.05$

Table 6.18 Significance Levels for Specific Menus by Age Group for Interrupt Rate

The within subject analysis indicated that both younger and older subjects interrupt significantly more often when using keypad entry. However, this increase in interrupt rate is present in all three component processes levels for the older subjects only.

Silence rates at each specific menu level were next to be analysed. Figure 6.5 shows the mean specific silence rates for the two groups of subjects.

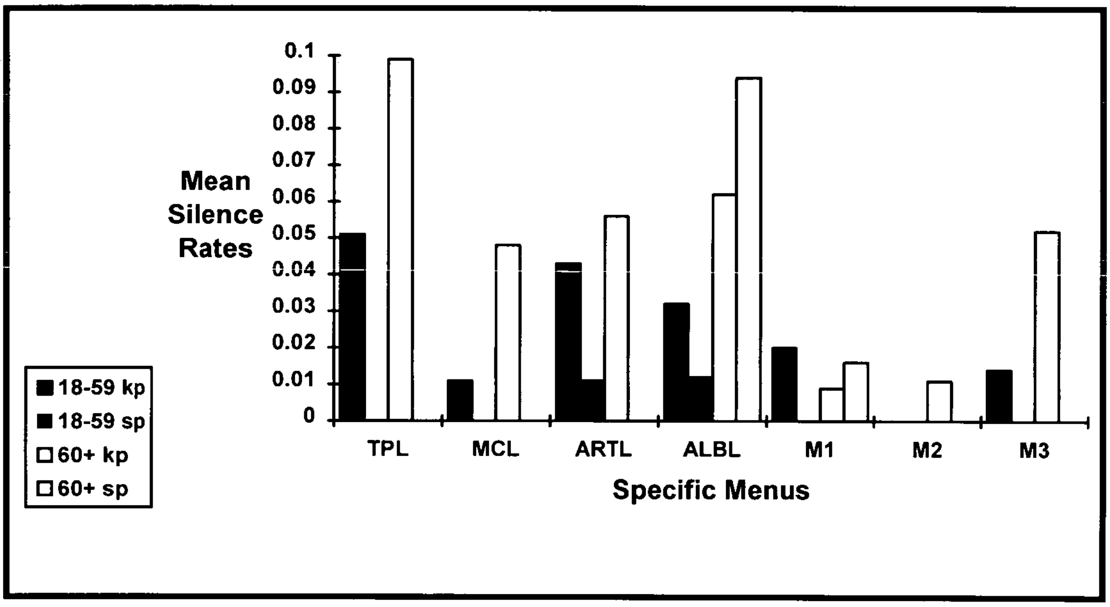


Figure 6.5 Mean Silence Rate for Specific Menus by Age

A within subject analysis indicated that older subjects were significantly less silent at the top level menu of the automated music catalogue when they were using speech input, {t = 3.33, df = 21, p < 0.01}. No significant within subject differences were observed for younger subjects at this menu level. A similar finding was observed for older subjects at the music category level, {t = 2.72, df =21, p < 0.05}, the artist level, {t= 2.84, df = 21, p < 0.01} and at the menu 3 level of the automated music catalogue, {t = 2.40, df = 21, p < 0.05}. Once again there were no corresponding results for younger subjects at these specific menu levels.

Individual Difference # 27

Older users have a significantly lower silence rate when using speech input.

The results presented here indicate that older subjects improved their performance, in terms of making fewer silence errors, when they used the speech input version of the automated music catalogue. This improvement in performance is present in the finding and navigation component process levels.

6.3.4.3. Summary of Component Process Analysis

The results of the component process analysis indicate that for interrupt both low and high verbal ability subjects interrupt more when using the keypad version of the automated music catalogue. A similar finding was also observed when age group differences are concerned. Thus it would appear that for interrupt rates at least, speech input does not readily accommodate older subjects and low verbal ability subjects. However, when silence rates were examined, a different picture emerges. Overall the silence rate results indicated that there were significant differences in silence rate between subjects when keypad entry was the input mode used. When the silence rates for speech input are examined, the results indicated that overall, the performance of the older subjects was matching that of the younger ones. These data indicate that speech was accommodating the older subjects when silence rate was taken a measure of performance. A within subject analysis of the older subjects at the component process level indicated that this improvement in performance was present at both levels (finding and navigation) which had been identified in the isolate analysis of Chapter 5.

6.3.5. Multiple Regression Analysis

Having shown where the significant differences occur for age and verbal ability at the component process level the next step was to use these two salient individual characteristics to predict performance on each component process level. Table 6.19 shows the standardised coefficients for age and verbal ability using mean interrupt rate as the criterion value for each component.

Input Mode	Component	Age	Verbal Ability	R^2
	finding	-0.331 ***	-0.001	0.261 **
Keypad	generating	-0.336 ***	0.005	0.348 ***
	navigation	-0.395 ***	-0.002	0.279 **
	finding	-0.449 ***	0.002	0.376 ***
Speech	generating	-0.366 ***	0.005	0.348 ***
	navigation	-0.395 ***	-0.002	0.279 **

Table 6.19 Relationships among Component Tasks and Age and Verbal Ability using Interrupt Rate as Criterion Measure

**p < 0.01, ***p < 0.001

A multiple regression analysis was then carried out using silence rates for each of the component process levels as the criterion values. Table 6.20 shows the standardised coefficients for age and verbal ability and how they relate to component task performance.

Mode of Input	Component	Age	Verbal Ability	R^2
	finding	0.026	-0.001	0.098
Keypad	generating	-0.008	-0.001*	0.163*
	navigation	-0.035	-0.000	0.053
	finding	0.015	-0.000	0.033
Speech	generating	0.005	-0.000	0.056
	navigation	-	-	-

Table 6.20 Relationships among Component Tasks and Age and Verbal Ability using Silence Rate as Criterion Measure

*$p < 0.05$

The results of the multiple regression analysis, taking interrupt rate as a measure of performance, indicate that age had a significant effect on all three component processes for the keypad and speech input versions of the automated music catalogue. The effects of verbal ability appear to be confined to the generating process part of the task for keypad input in terms of silence rate only.

On the basis of these results, it would appear that the working hypothesis put forward in the previous Chapter relating age and verbal ability to the automated music catalogue component operations needs to be revised. The effect of verbal ability appears to be confined to the generating process part of the task for keypad input in terms of silence rate only. In the previous experiment, the effects of verbal ability were found in the navigation part of the task only. These results indicate that it is difficult to significantly predict (isolate) where the effects of verbal ability in this type of task occur. Even where it was possible to isolate the effects of verbal ability the results only just reached significance ($p < 0.05$). In terms of age effects however, the results of the regression analysis (based on interrupt rate performance for both keypad and speech input modes) confirms the effects found for age in the previous experiment; age significantly effects interrupt rate performance in all three component processes - finding, generating and navigation. On the basis of these results a modified version of the model presented in the previous chapter of this thesis needs to be constructed. Figure 6.6 shows the revised working hypothesis for both the speech input and keypad versions of the automated music catalogue.

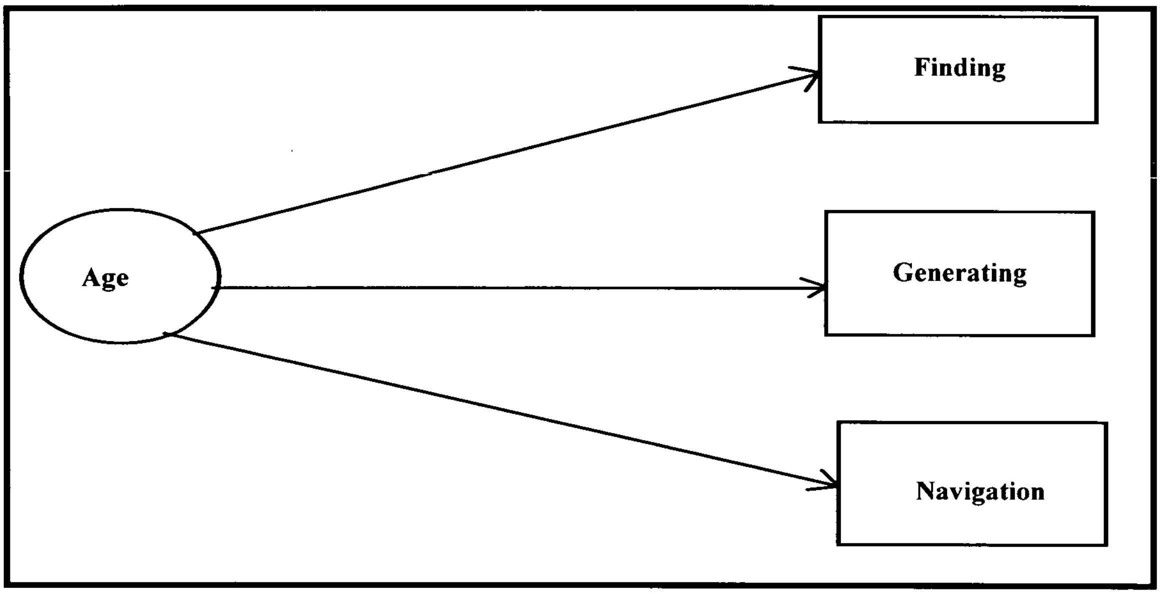


Figure 6.6 Working Hypothesis Relating Age to Automated Music Catalogue Component Operations for Speech and Keypad Input Modes

6.3.5. Conclusions

The results obtained from the experiment reported in this chapter suggest that the experimental hypothesis can only be partially accepted. There is a significant improvement in the performance of older subjects, in terms of a reduction in silence rate, when they use the speech input version of the automated music catalogue. Their performance here matches that of the younger group of users. Therefore it can be said that speech is accommodating them in this instance. However, when interrupt rates are considered for both the older subjects and those classed as having low verbal ability, it was found that these subjects interrupt more on the keypad version of the service, though not to the same extent as the high verbal ability and younger group of subjects. These results correspond to the finding obtained in the experiment reported in the previous Chapter of this thesis. These differences in the performance measures can be partially explained by the fact that older people (it emerged from the semi-structured interviews after the experiment was completed) are not as confident as young people in their use of new technology. Therefore they will not interact with the system in the same way as would younger, more experienced users of new technology (e.g. older subjects interrupt system prompts less often than younger users). Silence rates however can be classed as a form of error and therefore offer a more basic measure of performance.

In terms of predicting performance, the results of the multiple regression analysis indicate that age is a better predictor of performance for both the keypad and speech versions of the automated music catalogue than verbal ability.

In sum, it can be said that speech input partially accommodates the needs of older subjects when using the automated music catalogue and this improvement is evident at the finding and navigation components of a task of this kind.

Chapter 7 : Conclusions

The aim of this thesis was to offer a practical approach to the development of interactive spoken dialogue systems for automated telephone services. As previous research has shown, there are some people who are more successful users of interactive spoken dialogue systems than others. However, factors which lead to this division have not been satisfactorily explained. The experimental work reported in this thesis offers a three stage approach, based on the work of Egan and Gomez (1985), to explain these differences and provide information for the design of systems that are more effective and efficient for a larger number of people.

The first stage of this approach - ***assaying*** - is to assess if individual differences have an effect on user performance on a task which involves using an automated telephone system. This was the aim of the experiment reported in Chapter Four. The experiment focused on three different types of input mode for an automated catalogue service: connected word, isolated word and keypad, with users using all three input modes. The experimental task required subjects to order an item from a catalogue giving a customer identification number, a catalogue item number and finally a credit card number. The results showed differences in user performance (measured by call duration) between the three age groupings (18-38, 39-59 and 60+), low and high verbal ability subjects and low and high spatial ability subjects on the keypad and connected word modes of the automated catalogue service. In addition, a multiple regression analysis indicated that individual differences such as age and verbal ability could be used to predict user performance (again based on call duration) when subjects were using the keypad version of the automated catalogue service. No significant differences in performance were found between any of the individual differences measured when subjects were using the isolated word version of the service. There were also no significant differences found between high and low scores on the five measures of personality (neuroticism, extraversion, openness, agreeableness and conscientiousness) in respect of performance on any of the three input modes. No gender differences were observed in terms of performance on the three versions of the automated catalogue service as well. When it came to analysing the attitude scores, no significant differences were found between any of the individual characteristics which were chosen for study in this experiment. Overall, this experiment indicated that it was possible to identify (assay) salient individual differences in users' performance on a task as simple as ordering an item from the automated catalogue service.

The second stage of the Egan and Gomez approach - ***isolating*** - was aimed at identifying those parts of the task where the salient individual differences have their greatest effect when subjects are interacting with an automated telephone service. This was the aim of the experiment reported in Chapter Five. The experiment was also designed to take into account the effects of learning on performance and in order to achieve this, subjects were given a more complex task to complete. The task used in this experiment required subjects to order 4 tracks to put on a personalised CD from an automated music catalogue, using keypad input. Subjects were required to attend the laboratory at the University of Edinburgh four times, at one week intervals, to complete a version of this task. The automated music catalogue was a hierarchically structured database comprising five categories of music: Blues, Jazz, Classical, Rock & Pop and Folk. There were four artists available for each category of music. Each artist had three albums available in the music catalogue and three tracks were available for selection by subjects from each album. In addition, subjects could listen to a selection from each track for a period of up to thirty seconds. As well as looking to confirm the assay of the first experiment, other measures of individual differences (information processing ability and memory) were also assessed in the hope of providing further information to explain the variance in subject performance.

The results from this experiment showed that over the four experimental trials, age and verbal ability significantly contributed to the variance in subject performance on a task of this kind, with the younger and high verbal ability subjects performing better than the older and low verbal ability subjects. Neither information processing ability nor memory made a significant contribution to performance differences. Spatial ability, unlike the first experiment (reported in Chapter 4 of this thesis), had only a limited effect on performance. Personality did not make a significant contribution to the variance in performance, once again producing results similar to those found in the first experiment. This result shows that although personality is considered to be one of the most stable and important characteristics of the individual, it does not appear to have a significant effect on subject performance on a human-computer task of this kind. In addition, no significant differences were found between the performances of male and female subjects. Once again, no significant differences in attitude were observed for any of the individual characteristics highlighted in this experiment.

The CD task was then decomposed, on the basis of the Egan and Gomez methodology, into three processes: finding, generating and navigation. Finding can be defined, in this instance, as being the process whereby subjects find the correct category of music, the correct artist within that category of music, the correct album by the artist they are interested in and the correct track from that album. The top level, music category level, the artist level and the album level, are the specific levels in the database structure of the automated music catalogue which correspond to this process. Generating can be defined as the process which allows subjects to play, select or add to their personalised CD the specific track they have found. Menu 1 and Menu 2 are the specific menus in the automated music catalogue which correspond to the generating process. Finally, navigation refers to the process which allows subjects to move around the database structure of the automated music catalogue. For example, after selecting a track to add to their CD, subjects can then use Menu 3 to navigate their way to the top level of the database in order to select another music category from which they will eventually select another track to add to their CD.

Further analysis was carried out (by using a multiple regression technique) which indicated that age had a significant effect on all three component processes whereas the effects of verbal ability are confined to the generating process. Therefore the experiment showed that it was possible to identify salient individual differences in a automated music catalogue task of this kind and isolate where some of these differences have their greatest effect.

The final stage of this approach - ***accommodating*** - aims to accommodate those individual differences that have been highlighted as having a significant effect on user performance in the previous experiment. This was the aim of the experiment reported in Chapter 6 of this thesis. From the results obtained in the experiment reported in Chapter 5 it was decided to try and accommodate older and low verbal ability subjects in the automated music catalogue task by providing them with a speech input version of the service. This is an appropriate step to take as it emerged from interviews conducted at the end of the experiment reported in Chapter 5 that these subjects felt they could have performed better if they had been allowed to speak to the service instead of using keypad entry. The experiment was designed to give half of the subjects the keypad version of the service first and then the speech version while the other half of the subjects used the speech input version of the service first followed by the keypad version of the automated music catalogue. This balanced design would allow a direct comparison of performance between the two versions of the service. All of the subjects who took part in this experiment were classed as being naive users of automated telephone services.

The results obtained from this experiment indicate that there is a significant improvement in the performance of older subjects, in terms of a reduction in silence errors, when they use the speech input version of the automated music catalogue. The performance of these older subjects matches that of the younger subjects in this instance. Speech input therefore appears to be accommodating older subjects in this instance. There was no significant decrease in the silence rate for low verbal ability subjects in comparison to high verbal ability subjects. When interrupt rates are considered, a slightly different picture emerges. In this instance, both the older and low verbal ability subjects interrupt more often when they use the keypad version of the automated music catalogue in comparison to the speech input version - although, once again, these subjects did not interrupt at the same rate as the younger and high verbal ability subjects in either version of the system. Once again, no significant differences were found for age, verbal ability and gender in terms of attitude towards both the speech and keypad input versions of the automated music catalogue service. Overall, the results from this final experiment indicate that speech input was only partially successful in accommodating older and low verbal ability subjects.

The aim of this thesis was to identify individual characteristics which have a significant effect on users' performance and attitude when using automated telephone services and to investigate methods to accommodate these differences in order to improve the usability of the service for the variety of individuals who will use it. By adopting the three stage approach of Egan and Gomez the experimental work reported in this thesis has shown that it is possible to identify certain individual characteristics which have a significant effect on users performance with an automated telephone service as well as identifying those that do not. In addition, this framework has allowed an investigation of where these salient individual differences have their greatest effect in the users interaction with the service by breaking the task down into three component processes: finding, generating and navigation. Finally, the methodology of Egan and Gomez states that once significant individual differences have been identified and isolated, the next stage is to try and accommodate these by improving the usability of the system. In terms of the work reported in this thesis, the method chosen to accommodate the salient individual differences identified in the first two experiments can only claim to have been partially successful.

The results of the experimental work reported in this thesis have also produced several important findings in relation to the design of interfaces for automated telephone services. Firstly, it has been shown that designers, by looking for ways to interpret the results produced by the performance measures they have taken, have the opportunity to design interfaces for automated telephone services which will be effective and efficient to use for the majority of people who wish to use the service. Secondly, individual differences such as age and verbal ability have been shown to have a significant effect on users' interactions with automated telephone services. Thirdly, by adopting a methodology like the three stage approach advocated by

Egan and Gomez, the designer has the opportunity to ascertain where these salient individual characteristics have their greatest effect in users' interaction with the system. By adapting the Egan and Gomez methodology, which was developed with graphical user interfaces in mind, this thesis has shown that an individual's interaction with an automated telephone service, such as the music catalogue, can be broken down into a three stage process of finding, generating and navigation.

Having shown the validity of this methodology to assess individual differences for automated telephone services, the next step forward for this work should focus on ways of adequately accommodating the individual differences which have been identified in the experimental results. This could take the form of a longitudinal study which explores the effects of learning on users periodic use of an automated telephone service like the automated music catalogue. This will provide a situation approximating individual's use of certain automated telephone services in real life. It should be remembered that users are aware that the interaction is costing them money and this is an important consideration when it comes to individuals continuing to use an automated service. If they do not feel they are getting value for money, they will not use a service. Another design improvement for improving the usability of a system like the automated music catalogue would be the opportunity for users to request on-line help when they felt the need.

New developments in word-spotting technology will further serve to allow accommodation of the individual differences identified in this work, leading to the appearance of better automated telephone services for people and more efficient user interfaces in these products.

Appendix A:

Likert Statements

1. I found the Automated Music Catalogue was easy to use

2. I found the pace of the Automated Music Catalogue was too slow

3. I thought the Automated Music Catalogue was efficient

4. I had to concentrate hard when using the Automated Music Catalogue

5. I felt that the Automated Music Catalogue needs a lot of improvement

6. I felt flustered when using the Automated Music Catalogue

7. I felt hurried when using the Automated Music catalogue

8. I thought that the Automated Music Catalogue was reliable

9. It would have been useful if I could have requested more help from the Automated Music Catalogue

10. I felt I was in control while using the Automated Music Catalogue

11. I would prefer to speak to a human being when ordering a personalised CD

12. I found the wording of the messages too repetitive

13. I thought the Automated Music Catalogue was polite

14. I thought the Automated Music Catalogue was friendly

15. I liked the voice

16. I felt impatient when I was using the Automated Music Catalogue

17. I found it was easy to make my responses using the telephone keypad

18. I thought the Automated Music Catalogue was too complicated

19. It was not always obvious how to find what I wanted in the Automated Music Catalogue

20. I thought the way that the choices were presented was clear

21. I would be happy ordering a personalised CD using the Automated Music Catalogue

22. I thought that it took too long to order the four tracks

23. I thought that learning to use the Automated Music Catalogue was easy

24. I enjoyed using the Automated Music Catalogue

25. I felt under stress while using the Automated Music Catalogue

26. I think the Automated Music Catalogue is a good idea

27. Sometimes I felt I was lost while using the Automated Music Catalogue

28. I thought the voice was very clear

29. It was always clear which key I needed to press

30. I found the Automated Music Catalogue was confusing to use

31. I found the instructions I read before using the Automated Music Catalogue were helpful

32. I felt that sometimes I was given too many possibilities to choose from

Appendix B: Documentation Provided for Subjects in Experiments

Experiment 1:

Priming for Isolated Word Input:

- You are going to order <state item> using an automated catalogue service
- First of all, you will be asked for your customer identification number
- Secondly, you will be asked for the number of the catalogue item you wish to order
- Finally, you will be asked for your CCIR credit card number
- Please read out each number in blocks of four digits, clearly saying one digit after each beep you hear
- If you are asked any questions by the service, only answer with either yes or no

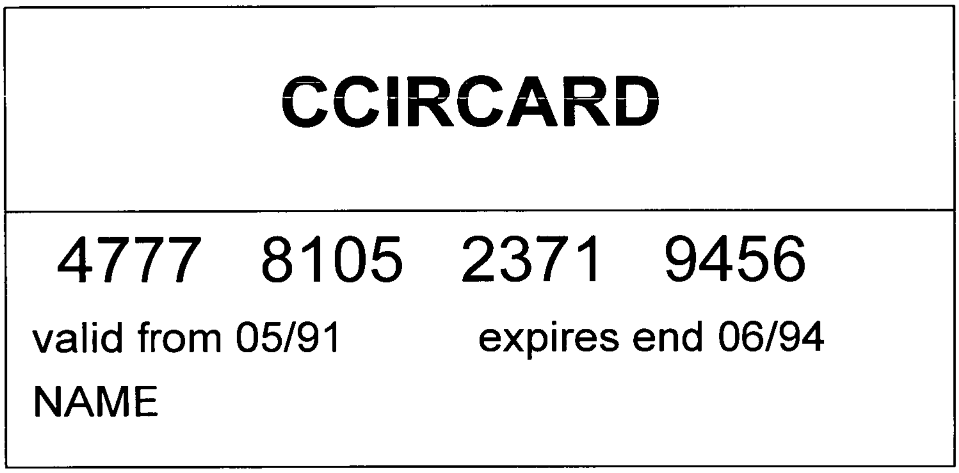
Priming For Connected Word Input:

- You are going to order <state item> using an automated catalogue service
- First of all, you will be asked for your customer identification number
- Secondly, you will be asked for the number of the catalogue item you wish to order
- Finally, you will be asked for your CCIR credit card number
- Please read out each number in blocks of four digits
- Please read out each block exactly as it is written, otherwise the service will not understand what you say
- If a block contains two ones together, for example, please say this as 'one, one' and not 'double one'
- If you are asked any questions by the service, only answer with either yes or no

Priming for Keypad Input:

- You are going to order <state item> using an automated catalogue service
- First of all, you will be asked for your customer identification number
- Secondly, you will be asked for the number of the catalogue item you wish to order
- Finally, you will be asked for your CCIR credit card number
- Please key the numbers in using the telephone keypad
- If you are asked any questions by the service, please respond by pressing the appropriate key on the keypad

Examples of credit card and customer identification number:

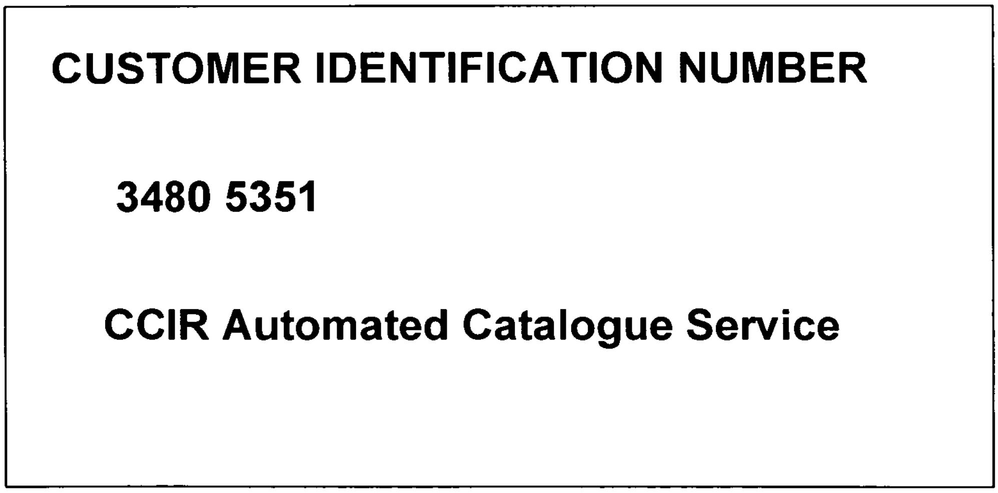


CCIRCARD

4777 8105 2371 9456

valid from 05/91 expires end 06/94

NAME

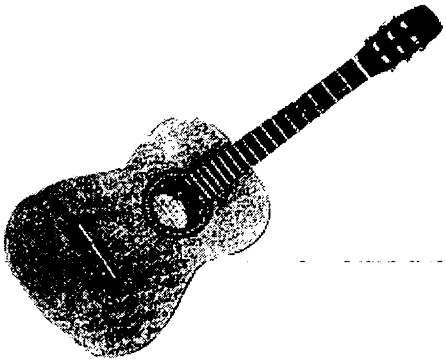


CUSTOMER IDENTIFICATION NUMBER

3480 5351

CCIR Automated Catalogue Service

Example Page from the CCIR Automated Catalogue:



Acoustic Guitar

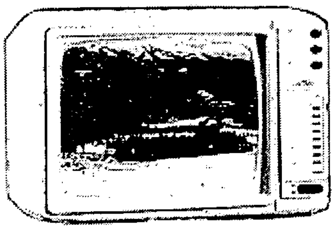
£35.99

Cat.No. 6384 5219

Amethyst Ring

£49.50

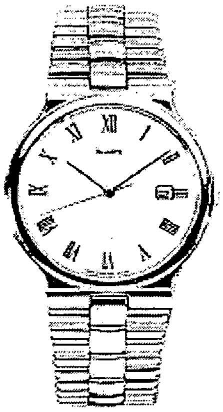
Cat.No. 9873 4156



Colour Television Set

£85.60

Cat.No. 7643 1525



Gold Watch

£38.45

Cat.No. 3225 1002

Experiment 2:

Priming for Automated Music Catalogue:

- The automated music catalogue lets you compile your own personalised CD by selecting tracks from a range of albums, artists and categories of music.

- To help you make your selection, the service lets you listen to short extracts of music of your choice. You can then add the tracks you like to your CD.

- At each point the service will offer you choices. Listen carefully and respond by pressing ONE of the keys on your telephone keypad. If you make a mistake or do not press a key the service will ask you to try again. If you know which key you want to press, you can press it before the voice finishes.

- At any time you can press the * (Star) key to obtain more options. These include listing the music categories again or leaving the catalogue.

What You Have to DO:

Please select the following four tracks for your personalised CD:

A. **Folk Category:**
from the Joni Mitchell album 'Blue' select the track "California"

B. **Classical Category:**
from the Brahms album 'Number 1 Intermezzo' select the track "Adagio"

C. **Jazz Category:**
listen to all the tracks on the John Coltrane album 'Blue Train' and then select one for your CD

D. **Jazz Category:**
listen to all the tracks on the Art Pepper album 'No Limit' and then select one for your CD

Experiment 3:

Priming for Automated Music Catalogue Voice Version:

- The automated music catalogue lets you compile your own personalised CD by selecting tracks from a range of albums, artists and categories of music.

- To help you make your selection, the service lets you listen to short extracts of music of your own choice. You can then add the tracks you like to your CD.

- At each point the service will offer you choices. Listen carefully and respond by saying the name of the item you want to select. If you make a mistake or do not say anything the service will ask you to try again. If you know which item you want to select, you can say the name of item before the voice finishes.

- At any time you can say the word "options" to obtain more information from the service. This information includes listing the music categories and leaving the catalogue.

What You Have to DO:

Please select the following <u>four</u> tracks for your personalised CD:

A. **Blues Category:**
from the Howlin' Wolf album 'The Wolf is at Your Door' select the track "Blue Bird"

B. **Classical Category:**
from the Brahms album 'Symphony Number 1' select the track "Allegro"

C. **Folk Category:**
listen to all the tracks on the Bob Dylan album 'Good As I Been to You' and then select one for your CD

D. **Folk Category:**
listen to all the tracks on the Joni Mitchell album 'Blue' and then select one for your CD

Priming for Automated Music Catalogue Keypad Version:

- The automated music catalogue lets you compile your own personalised CD by selecting tracks from a range of albums, artists and categories of music.

- To help you make your selection, the service lets you listen to short extracts of music of your choice. You can then add the tracks you like to your CD.

- At each point the service will offer you choices. Listen carefully and respond by pressing ONE of the keys on your telephone keypad. If you make a mistake or do not press a key the service will ask you to try again. If you know which key you want to press, you can press it before the voice finishes.

- At any time you can press the * (Star) key to obtain more options. These include listing the music categories again or leaving the catalogue.

What You Have to DO:

Please select the following **four** tracks for your personalised CD:

A. **Jazz Category:**
from the Chet Baker album 'Line for Lyons' select the track "My Funny Valentine"

B. **Folk Category:**
from the Bob Dylan album 'Bringing it All Back Home' select the track "Subterranean Homesick Blues"

C. **Classical Category:**
listen to all the tracks on the Bach album 'Cello Suite Number 1' and then select one for your CD

D. **Classical Category:**
listen to all the tracks on the Mozart album 'Symphony Number 40' and then select one for your CD

REFERENCES

Allport, G.W. & Oddbert, H.S. (1936) Trait-names : *A Psycho-lexical Study,* cited in Carver, C.S. & Scheier, M. F. (1992) *Perspectives on Personality,* Allyn and Bacon.

Atkinson, R.L., Atkinson, R.C. & Hilgard, E.R (1983) *Introduction to Psychology*, Harcourt Brace Jovanovich.

Atkinson, R.C. & Shiffrin, R.M. (1968) *Human memory: A proposed system and its control processes,* In K.W. Spence & J.T. Spence (Eds), The Psychology of Learning and Motivation, Vol. 2, London, Academic Press

Baber, C., (1993a) *Developing Interactive Speech Technology* in Interactive Speech Technology, Barber, C. & Noyes, J. (eds), Taylor & Francis

Baddeley, A.D. & Hitch, G. (1974) *Working Memory.* In G.H. Bower (Ed) The Psychology of Learning and Motivation, Vol. 8, London, Academic Press

Baddeley, A.D. , Thomson, N. & Buchanan, M. (1975) *Word length and the structure of short term memory,* Journal of Verbal Learning and Verbal Behaviour, 14, 575-589

Baddeley, A.D. & Lieberman, K. (1980) *Spatial working memory.* In R. Knickerson (Ed) Attention and performance, vol. 8, Hillsdale, NJ, Lawerence Erlbaum Associates Inc

Baddeley, A.D. (1986) *Working Memory,* Oxford, Oxford University Press

Baum, F. (1900) *The Wizard of Oz,* Collins. London (1974)

Benyon, D. (1993) *Accommodating Individual Differences through an Adaptive User Interface,* Adaptive User Interfaces, 149-165

Borgman, C.L. (1986) *"Why Are On-Line Catalogues Hard to Use? Lessons Learned from Information-Retrieval Studies"*, Journal of the American Society for Information Science, 37 (6), 387-400.

Botwin, M.D. & Buss, D.M. (1989) *"Structure of Act-report Data : Is the Five-Factor Model of Personality Recaptured?"*, Journal of Personality and Social Psychology, 56, 988-1001.

Borgotta, E.F. (1964) *"The Structure of Personality Characteristics"*, Behavioural Science, 12, 8-17.

Card, S.K, Moran, T.P & Newell, A. (1984) The Psychology of Human Computer Interaction, Hillsdale, NJ: Lawrence Erlbaum Associates.

Carroll, J.B. (1983) *Studying Individual Differences in Cognitive Abilities: Through and Beyond Factor Analysis.* In Dillon, F.R. & Schmeck, R.R. (1983) Individual Differences in Cognition Volume 2, Academic Press, London.

Carver, C.S. & Scheier, M. F. (1992) *Perspectives on Personality,* Allyn and Bacon.

Cattell, R.B. (1946) *The Scientific Analysis of Personality and Motivation,* New York Academic Press.

Cattell, R.B., Eber, H.W. & Tatsuoka, M.M. (1970) *The British Standardisation of the 16PF.*

Cattell, R.B. & Kline, P. (1977) *The Scientific Analysis of Personality and Motivation,* New York: Academic Press.

Chapanis, A. (1981) *Interactive Human Communication: Some Lessons Learned from Laboratory Experiments,* in Man-Computer Interaction: Human Factors Aspects of Computers and People (Shakel, B. ed) Sijhoff and Noordhoff, Rockville, Maryland.

Choinere, A., Robert, J.M. & Descout, R. (1991) *Building a User Interface for a Speech-Based Telephone Application System,* Proceedings of EuroSpeech 1991, Volume 3, pp1503-1506.

Coolican, H. (1994) *Research Methods and Statistics in Psychology,* Hodder & Stoughton

Cooper, L.A. & Mumaw, R.J. (1985) *Individual Differences in Cognition Volume 2,* Academic Press, London.

Costa, P.T., Jr, & McCrae, R.R. (1986) *Major Contributions to Personality Pschology* in Modgil, S. & Modgil, C. (eds) *Hans Eysenck: Concensus and Controversy* Philadelphia: Falmer

Costa, P.T., Jr, & McCrae, R.R. (1987) *Neuroticism, Somatic Complaints and Disease: Is the Bark Worse than the Bite?,* Journal of Personality and Social Psychology, 55, 299-316.

Costa, P.T., Jr, McCrae, R.R. & Dye, D.A. (1991) *Facet Scales for Agreeableness and Conscientiousness: A Revision of the NEO Personality Inventory,* Personality and Individual Differences, 12, 887-898.

Costa, P.T., Jr, & McCrae, R.R. (1992) *Normal Personality Assessment in Clinical Practice: The NEO Personality Inventory,* Psychological Assessment, 4, 5-13, 20-22.

Cox. G. & Cooper, D. (1981) *Speech Communication and How to use it,* in Fundamentals of HCI, edited by Andrew Monk, Lawrence Earlbaum Associates.

Cronbach, L.J. & Snow, R.E. (1977) *Aptitudes and Instructional Methods,* New York: JohnWiley & Sons

Damper, R.I. (1993), *Speech as an Interface Medium: How Best can it be Used,* in Interactive Speech Technology, Barber, C. & Noyes, J. (eds), Taylor & Francis

Digman, J.M. (1990) "Personality Structure: Emergence of the five-factor model", *Annual Review of Psychology.*

Digman, J. M. & Inouye, J. (1986) *"Further specification of the five robust factors of Personality"*,Journal of Personality and Social Psychology, 50, 116-123.

Dutoit, D. (1987) *"Evaluation of Speaker-Independent Isolated-Word Recognition Systems over the Telephone Network",* Proceedings of EuroSpeech 1987, 241-244.

Egan, D. (1988) *"Individual Differences in Human-Computer Interaction",* Handbook of Human-Computer Interaction, Elsvier Science.

Entwhistle, N.J. (1978) *Knowledge of Structures and Styles of learning: A Summary of Pask's Recent Research,* British Journal of Educational Psychology, 48, pp. 255-265.

Egan, D.E. & Gomez, L.M. (1985) *"Assaying, Isolating and Accommodating Individual Differences in learning a Complex Skill"*, Individual Differences in Cognition, Volume 2, Academic Press.

Ekstrom, R.B., French, J.W. & Harman, H.H. (1976) *Manual Kit for Factor- Referenced Cognitive Tests,* Princeton, NJ: Educational Testing Service.

Elkerton, J. & Williges, R.C. (1984) *"Information Retrieval Strategies in File-Search Environment",* Human Factors, 26 (2) 171-184.

Eysenck, H.J. (1975) *The Scientific Study of Personality,* MacMillan.

Eysenck, M. W. & Keane, M.T. *Cognitive Psychology: A Students Handbook,* Lawrence Erlbaum Associates.

Fay, D. (1993) *"Interfaces to Automated Telephone Services: Do Users Prefer Touch-Tone or Automatic Speech Recognition?"* Proceedings of Human Factors in Telecommunications, 1993, 339-349.

Fiske, D.W. (1949) *"Consistency of the factorial structures of personality ratings from different sources"*, Journal of Abnormal and Social Psychology, 44, 329-344.

Foster, J.C., Dutton, R., Love, S., Nairn, I.A., Vergeynst, N & Stentiford, F.W.M. (1993) *Intelligent Dialogues in Automated Telephone Services,* in Interactive Speech Technology: Human Factors Issues in the Application of Speech Input/Output to Computers (Barber, C. & Noyes, J.N., eds.) Taylor & Francis.

Fraser, N.M. & Gilbert, G.N. (1991) *Simulating Speech Systems,* Computer Speech and Language, 5, 81-99.

Fraser, N.M. (1995) *"Quality Standards for Spoken Language Dialogue Systems: A Report on the Progress in EAGLES"*, Proceddings of the ESCA Workshop on Spoken Language Dialogue Systems,Vogso, 1995

Goldberg, L.R. (1981) *"Language and Individual Differences: The search for universals in personality lexicons"* in Perspectives on Personality, (Ed) Carver, C.S. & Scheier, M.F., Allyn & Bacon.

Gomez, L.M., Egan, D.E. & Bowers, C. (1986) *"Learning to use a text editor: some learner characteristics that predict success"*, Human-Computer Interaction 2, 1-23.

Gould, J.D. and Boies, S.J. (1984) *"Speech Filing - an Office System for Principles,* cited in Martin, M.M., Williges", B.H. & Williges, R.C. *"Improving the Design of Telephone-Based Information Systems"*, Proceedings of the Human Factors Society 34th annual meeting, 1990, 198-202.

Gronwall, D. (1977) *Paced Auditory Serial Addition Task: A Measure of Recovery from Concussion,* Perceptual and Motor Skills, 44, 367-373.

Greene, S.L., Gomez, L.M. & Devlin, S.J. (1986) *"A cognitive analysis of of database query production"*, Proceedings of the Human Factors Society,9-13.

Guyomard, M. & Siroux, J. (1988) *Constitution incrementale d'un Corpus de Dialogues Oraux Cooperatifs,* Journal Acoustique, 1, 329-337

Halstead-Nusslock, R. (1989) *"The Design of Phone-Based Interfaces for Consumers"* Proceedings of CHI' 1989, 347-352.

Hansen, W.J. (1971) *"User Engineering Principles for Interactive Systems",* Proceddings of the Fall Joint Computer Conference, 39, AFIPS Press, 523-532.

Hauptman, A.G. & Rudnicky, A.I. (1988) *Talking to Computers: An Empirical Investigation,* International Journal of Man-Machine Studies, 28, 583-604

Hauptman, A. G. (1989) *Speech and Gestures for Graphic Image Manipulation,* in CHI '89, 241-245.

Heim, A.W., Wallace, J.G. & Cane, V.R. *The effects of repeatedly retesting the same group on the same intelligence test. I Normal Adults, II High-Grade Mental Defectives, III Further Experiments and General Conclusions.* Quarterly Journal of Experimental Psychology, 1949.

Heim, A.W. (1970) *AH Series,* NFER-NELSON.

Hitch, G.J. (1978) *The role of short-term working memory in mental arithmetic.* Cognitive Psychology, 10, 302-323

Hitch, G.J. & Baddeley, A.D. (1976) *Verbal reasoning and working memory.* Quarterly Journal of Experimental Psychology, 28, 603-621.

Hunt, E. B. , Frost, N., & Lunneborg, C. (1973) *Individual Differences in Cognition: A New Approach to Intelligence,* in G. Bower (ed) The Psychology of Learning and Motivation (vol. 7) New York: Academic Press.

Hunt, E. & Lunneborg, C. & Lewis, J (1985) *What does it mean to be high verbal,* Cognitive Psychology, 8, 194-227.

Jennings, F., Benyon, D.R. & Murray, D.M. (1991) *Adapting Systems to Differences Between Individuals,* in Acta Psychologica 78, nos. 1-3, 248-258.

Jennings, F. & Benyon, D. (1989) "Databases: Different Interfaces for Different Users", *NPL Report.*

Johnson, J.A. (1981) *The "Self-Disclosure and "Self-Presentation views of Item Response Dynamics and Personality Scale Validity,* Journal of Personality and Social Psychology, 40, 761-769.

Jones, D., Hapeshi, K. & Frankish, C. (1989) *Design Guidelines for Speech Recognition Interfaces,* Applied Ergonomics, 1989, 20.1., 47-52.

Jung, C. (1971) *Psychological Types,* Kegan Paul Harcourt Brace & Co.

Karis, D. & Dobroth, K. (1991) *"Automating services with speech recognition over the public switched telephone network: human factors considerations"* IEEE Journal on Selected Areas in Communications, Vol. 9, No. 4, 1991, 574-585.

Koubek, R.J., LeBould, W.K. & Salvendy, G. (1985) *"Predicting performance in computer programming courses"*, Behaviour & Information Technology, 4 (2) 13-129.

Labrador, C. & Dinesh, P. (1984) *Experiments in Speech Interaction with Conventional Data Services,* in Interact '84, 104-108

Lea, W.A., (1980), *The Value of Speech Recognition Systems*, in Trends in Speech Recognition, W.A. Lea (ed), Prentice-Hall

Lezack, M.D. (1983) *Neuropsychological Assessment,* 2nd edition, 422-429, Oxford University Press: Oxford.

Likert, R.A. (1932) *A Technique for the Measurement of Attitudes,* Archives of Psychology, 140, 55.

Lohman, D.F. (1979) *Spatial Ability: A review and reanalysis of the correlational literature,* (Technical Report number 8), Aptitude Research Project, School of Education, Stanford University.

Martin, M.M., Williges, B.H. & Williges, R.C. (1990) *Improving the Design of Telephone-Based Information Systems*, Proceedings of the Human Factors Society 34th Annual Meeting, 1990, pp198-202.

McCrae, R.R. & Costa, P.T. (1990) *Personality in Adulthood,* The Guilford Press.

McCrae, R.R. & Costa, P.T. (1987) *"Validation of the five-factor model of personality across instruments and observers",* Journal of Personality and Social Psychology, 52, 81-90.

McFarlane, M.A. (1925) *A study of practical ability,* British Journal of Psychology Monograph Supplement, 8.

McGee, M.G. (1979) *"Human spatial abilities: Psychometric studies and environmental, genetic, hormonal, and neurological influences"* Psychological Bulletin, 86, 889-918.

Morgan, K. & MacLeod, H. (1990) *"The possible role of personality factors in computer interface preference"*, in Second Interdisciplinary Workshop on Mental Models, Robinson College, Cambridge.

Murray, D. & Bevan, N. (1984) *"The Social Psychology of Computer Conversation",* INTERACT '84, Elsvier Science.

Murray, A.C., Frankish, C.R. & Jones, D.M. (1991) *"System Design and Human Factors in Auditory Interfaces",* Proceedings EuroSpeech 1991, 1507-1510.

Myres, M.B. & McCaulley, M.H. (1985) *"The Myres-Briggs Type Indicator"*, Saville-Holdsworth.

Newell, A.F. (1987) *Speech Simulation Studies - Performance and Dialogue Specification,* in Recent Developments and Applications of Natural Language Understanding, Unicom Seminar, December.

Nickerson, R.S. (1976) *"On Conversational Interaction with Computers, User Orientated Design of Graphic Systems"*, ACM.

Norman, W.T. (1963) *"Towards an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings",* Journal of Abnormal and Social Psychology.

Noyes, J. (1993), *Speech Technology in the Future,* in Interactive Speech Technology, Barber, C. & Noyes, J. (eds), Taylor & Francis

Pask, G. (1976) *Styles and Strategies of Learning,* British Journal of Educational Psychology, 46, 12-25.

Pask, G. (1980) *Developments in Conversation Theory - Part 1,* International Journal of Man-Machine Studies, 13, pp. 357-411.

Peckham, J. (1993) *"A New Generation of SpoeknLanguage Dialogue Systems: REsults and Lessons from the SUNDIAL Project",* Proceedings of EuroSpeech 1993, 33-40.

Peckham, J. (1995) *"Conversational Interaction: Breaking the Usability Barrier",* ESCA Workshop on Spoekn Language Systems, Visgo, 1995, 1-8.

Poulson, D. (1987) *Towards Simple Indicies of the Perceived Quality of Software Interfaces,* in IEE Colloquium - Evaluation Techniques for Interactive System Design, IEE, Savoy Place, London.

Raven, J.C. (1958) *Standard Progressive Matricies,* H.K. Lewis & Co.

Richards, M.A. & Underwood, K.M. (1984a) *"Talking to Machines. How are People Naturally Inclined to Speak?"* Contemporary Ergonomics (Megaw, E.D., ed) Taylor and Francis, London.

Roberts, T.L. & Moran, T.P. (1983) *"The Evaluation of Text Editors: Methodology and Experimental Results",* Communications of the ACM, 26 265-283.

Rossen, M.B. (1983) *"Patterns of experience in text editing",* Proceedings of the CHI'83 Human Factors in Computing Systems, 171-175.

Rust, J.R & Golombok, S. (1989) *Modern Psychometrics: The Science of Psychological Assessment,* Routledge: London and New York.

Schmandt, C. (1987) *"Conversational Telecommunications Environments",* in Cognitive Engineering in the Design of Human-Computer Interaction and Expert Systems (ed, Salvendy, G.) Elsevier Science Publishers B.V., 1987.

Schneiderman, B. (1980) *Designing the User Interface,* Addison-Wesley Publishing Company.

Schneiderman, B. (1987) *Designing the User Interface: Strategies for Effective Human-Computer Interaction,* Addison-Wesley

Shock, N.W., Greulich, R.C., Anres, R, Arenberg, D., Costa, P.T., Jr, Lakatta, E.G. & Tobin, J.D. (1984) *Normal Human Aging: The Baltimore Longitudinal Study of Aging,* cited in Costa, P.T., Jr, & McCrae, R.R. (1992) NEO-PIR Professional Manual, Pschological Assessment Resources Inc.

Simpson, C.A., McCauley, M.E., Roland, E.F. Ruth, J.C. & Willges, B.H. (1985) *"System Design for Speech Recognition and Generation",* Human Factors, 27, 115-141.

Smith, G.M. (1967) *"Usefulness of peer ratings of personality in educational research",* in Perspectives on Personality, (ed. Carver, C.S & Scheier, M.F.) Allyn & Bacon, 1967.

Sternberg, R.J. (1977) *Intelligence, Information Processing and Analogical Reasoning: The Componential Analysis of Human Abilities,* Hiilsdale, NJ: Lawerence Earlbaum Associates.

Thurstone, L.L. (1938) *Primary Mental Abilities,* Psychometric Monographs.

Turkle, S. (1984) *The Second Self: Computers and the Human Spirit,* Granada Publishing.

Van der Veer, G.C. (1990) *Human-Computer Interaction: Learning, Individual Differences and Design Recommendations*, Offsetdrukkerij, Haveka, B.V.

Van der Veer, G.C, Tauber, M.J., Waern, Y. & van Muylwijk, B. (1985) *"On the Interaction Between System and User Characteristics"*, Behaviour and Information Technology, 4 289-308.

Vincente, K.J., Hayes, B.C. & Williges, R.C. (1987) *"Assaying, Accommodating and Isolating Individual Differences in Searching a Hierarchical System"*, Human Factors, 29 (3) 349-359.

Waern, Y. (1989) *Cognitive Aspects of Computer Supported Tasks*, John Wiley and Sons.

Warrington, E.K. & Shallice, T. (1972) *"Neuropsychological evidence of visual storage in short-term memory tasks".* Quarterly Journal of Experimental Psychology, 24, 30-40.

Waugh, N.C. & Norman, D. (1965) *Primary memory.* Psychological Review, 72, 89-104.

Witkin, H.A. & Goodenough, D.R. (1981) *Cognitive Styles: Essence and Origin,* International University Press.

Wiggins, J.S. (1979) *"A psychological taxonomy of trait-descriptive terms: The interpersonal domain. ",* Journal of Personality and Social Psychology, 37, 395-412.

Wilpon, J.G. (1989) *"Automatic Speech Recognition Over the DDD Telephone Network - Results of Extensive Network Field Studies",* Proceedings of SpeechTech '89, 133-136.