System-Initiated Digressions and Hidden Menu Options in Automated Spoken Dialogue Systems

THE UNIVERSITY OF EDINBURGH

Jenny G M Wilkie

PhD Thesis

2005

Declaration of originality

16th of December, 2005

This thesis is submitted in partial fulfilment for the degree of Doctor of Philosophy. I declare that it has been composed by myself, and that the work described is my own research.

Jenny G M Wilkie

Acknowledgements

I want to thank Professor Mervyn Jack, my friends and colleagues at the Centre for Communication Interface Research at the University of Edinburgh. I owe special thanks to my supervisor Dr John Foster for all his support and invaluable feedback on my work. Special thanks also go to Dr Fergus McInnes who took on the responsibility of supervisor following the retirement of Dr Foster and who, with admirable patience and a good sense of humour, introduced me to the weird and wonderful world of that which we call statistics. I also want to thank Peter Littlewood and colleagues at Lloyds TSB.

When the going gets tough, the tough go for coffee; thank you Hazel for being there to share my ups and downs over a cuppa, for listening, and for your support. I also want to thank my friends in Edinburgh, Rit, Nancie, Rory, Morven, David P, Lorna, Ian, Diarmid and Helen for all the laughs and good memories. I cherish my time spent in Scotland.

Finally, I want to thank my family (both in Scotland and in Sweden), my mum, my brother, and especially my husband David whom I admittedly adore. Special thanks go to my dad, a diesel engineer of the old school, who broadened my worldly views by sharing with me his enthusiasm for aerodynamics, pistons and DIY. Sadly, he passed away before the completion of my PhD project and I dedicate this thesis in his memory.

The current research was made possible by the generous support of Lloyds TSB Group, U.K. and Nuance Communications Inc.

To Dad

Abstract

Automated speech recognition technology is increasingly used in the mass-market domain of self-service telephone applications. However, the recognition technology is rarely infallible and system prompts in the form of explicit instructions, menu listings and telephone keypad touch-button options are commonly used in order to support the spoken human-computer interaction. A dilemma facing designers of such menu-based applications is where to place new or less frequently requested service options within the call-flow and how to incorporate these with the existing dialogue interface design. To date, this topic within the field of dialogue engineering has not been fully addressed.

The research detailed in this thesis proposes the use of system-initiated digressions as an alternative strategy to that of explicitly adding all options to the main menu listing of a speech-driven automated service. The purpose of these digressions was to deliver information about the availability of a new product or service option that could be triggered by using the relevant spoken keyword at the main menu. The keyword itself, however, was not explicitly mentioned (i.e. remained 'hidden') as an option in the existing main menu listing; therefore, callers had to infer that the option was available and then initiate the request themselves rather than passively select it from the menu listing.

The dialogue engineering investigation presented here centres on three main themes: the location of the digression in the dialogue, the turn-taking strategy employed and the type of register (wording) adopted. Contrasting system-initiated digressions were introduced into the dialogue of an existing real-world automated telephone banking service. In a series of four progressive empirical experiments, participants were invited to use the automated service to carry out banking tasks and were subjected to digressive dialogues in the form of banking product offers. The purpose of the experiments was to evaluate the impact of deploying system-initiated digressions on user attitudes toward the usability of the core service. Furthermore, detailed analyses were also performed to determine the effect that varying location, strategy and register in the digressions may have on participant attitudes. In particular, the thesis provides a novel approach to dialogue engineering by introducing established politeness theory in human-human interaction into the field of human-computer dialogues.

The conclusions drawn from this research support the introduction of system-initiated digressions in automated services. However, issues regarding users' mental models of menu-driven automated services and their expectations of the computer's social behaviour were identified in the research: participants had difficulties with correctly interpreting the concept of 'hidden' menu options and were sensitive to the more forceful registers adopted in the system-initiated digressions. The findings from the four experiments are presented and their impact on the development of future dialogue engineering strategies is discussed.

List of figures

List of figures

List of publications

Wilkie, J., McInnes, F., Jack, M.A., Littlewood, P., "Hidden menu options in automated human-computer telephone dialogues: dissonance in the user's mental model", *Behaviour and Information Technology* (in press).

Wilkie, J., Jack, M. A., Littlewood, P., (2005). "System-Initiated Digressive Proposals in Automated Human-Computer Dialogues: the Use of Contrasting Politeness Strategies", *International Journal of Human Computer Studies 62 (2005), pp. 41-71.*

Wilkie, J., Jack, M. A., Littlewood, P., (2002). "Design of System-Initiated Digressive Proposals for Automated Banking Dialogues", *Proceedings of ICSLP'02, Denver, Colorado, pp. 1493-1496.*

Abbreviations

DTMF: Dual Tone Multi Frequency (telephone keypad push-button input)

ETSI: European Telecommunications Standards Institute

GUI: Graphical User Interface

ISO: International Organisation for Standardisation

SDS: Spoken Dialogue System

SID: System-Initiated Digression

SLM: Statistical Language Modelling

VUI: Voice User Interface

Glossary

anaphora: in linguistics, the use of e.g. a pronoun to refer back to another unit instead of repeating a word

anthropomorphism: the ascription of human-like attributes and characteristics to an otherwise non-human object

digression: a part of a discourse not upon the main subject

dissonance: inconsistencies between the beliefs one holds, conflict between opinions or actions

interlocutor: someone who takes part in a conversation

menu: in an automated telephone service, a menu offers a user a list of choices from which a selection can be made

prompt: a system message (audio) which instructs the user about the kind of input that is expected

register: a variety of language used in a specific social setting

valence: the attraction or aversion that an individual feels toward a specific object or event

Contents

RESEARCH METHODOLOGY

EXPERIMENT 1 – STRATEGY FOR DELIVERY (PILOT STUDY)

EXPERIMENT 2 – DIALOGUE LOCATION OF SYSTEM-INITIATED DIGRESSIVE PROPOSALS

EXPERIMENT 4 – THE USE OF CONTRASTING POLITENESS STRATEGIES FOR SYSTEM-INITIATED DIGRESSIVE PROPOSALS

GENERAL DISCUSSION AND CONCLUSIONS

Chapter 1

Be a craftsman in speech that thou mayest be strong, for the strength of one is the tongue, and speech is mightier than all fighting.

- Maxims of Ptahhotep (~2300 BC), Egyptian Vizier -

Introduction

1.1 Introduction

The thesis expounded in this work is that system-initiated digressions can be included as a means for successfully introducing new products and services into the existing dialogue of a speech-enabled automated telephone service.

Speech plays an important part in the everyday communication between humans. Once mastered, it provides a flexible means of expression. It is portable, instantaneous and rich in information about personal characteristics and emotional states. The human ear and brain are finely attuned to perceiving speech input and, based on the interpretation of what is being said, it is possible to adjust one's own speech according to the perceived capabilities of the interlocutors: speaking at a slower rate, or louder, changing pitch, whispering, choosing different wordings and enunciating more carefully.

It comes as no surprise that speech also has attracted attention in the context of human-computer interaction, resulting in a substantial research effort where the aim is to enable machines to communicate verbally with human users. For decades, multidisciplinary teams consisting of artificial intelligence scientists, linguists, psychologists and signal processing engineers have worked on computational models for automating processes of speech production and perception. However, judging by the relatively limited range of readily available commercial applications on the market, equipping computers with the capability of understanding and generating human speech is not a straightforward feat. True conversational machines are – for the foreseeable future – the subject of science fiction novels and films.

There is one form of 'conversational' system that is increasingly, and successfully, being introduced in the mass-market: Spoken Dialogue Systems (SDSs) in the domain of self-service telephone applications. The development of such SDS applications has been spurred on by advances in speech recognition technology over the past 10-20 years, which has resulted in commercially available speech recognisers that allow robust, speaker-independent, continuous-word recognition with barge-in capabilities (enabling users to interrupt the computer output message). Still, the automated recognition process is far from infallible and therefore much of the dialogue design for mass-market automated telephone services is centred around making the interaction fit the technology at hand, relying on explicit (or implicit) instructions in order to guide users as to what to say, how to say it and when to speak; and to provide back-up error recovery strategies where the interaction between the human and the machine breaks down.

1.2 Automated telephone services and their limitations

Interactive Voice Response (IVR) applications in the form of touch-button operated telephone services have been used for decades to route customer calls and to provide direct access to information over the telephone. Speech recognition technology is increasingly being used to replace, or complement, the use of keypad input and now features in a number of mass-market applications such as cinema bookings, banking, travel information and e-mail over the phone.

Compared to their push-button counterparts, applications which use spoken language input offer users a more natural and flexible way of interacting with a computer-based system. However, the system messages and the turn-taking in these speech operated applications often still resemble those found in push-button operated services in that they follow a rigid prompt-response sequence where the input options are presented to users in the form of vocal menus and explicit instructions about what to say. The dialogue between the human user and the automated service in such applications typically follows a pre-defined script involving a fixed turn-taking structure (the computer prompts then the user responds) and valid user responses are restricted by the capabilities of the speech recognition grammar. Users of these mass-market applications can expect a controlled and predictable interaction with the computer in a dialogue that does not change between phone calls.

In mass-market applications, menus in the form of explicit list selections are usually employed as a method for informing users (especially novice users and in 'walk-up-and-use' systems) about the range of services available to them. Touch-tone key mappings for menu options are also often provided as an alternative input strategy alongside voice, for example, when the user may prefer to push telephone buttons (e.g. giving a sense of privacy when entering bank account information) or when the human-computer interaction needs supporting in order to avoid a breakdown (e.g. as fall-back after repeated mis-recognitions, or in noisy environments). The inclusion of touch-tone key mappings, coupled with the fact that system prompts frequently consist of pre-recorded human speech, lead to rather rigid application structures where changes once the service dialogue has been implemented and launched are impractical and can be costly. A dilemma facing designers of such menu-based applications is where to place new or less frequently requested service options within the call-flow and how to incorporate these with the existing interface design. Voice recognition design guidelines described in the recent literature (detailed in Chapter 2.6 of this thesis) offer broad coverage for how to implement menus in mass-market automated telephone services. However, they do not fully address issues surrounding the maintenance and future development of menu-driven services after they have been deployed.

1.3 System-initiated digressions and hidden menu options

There are a number of reasons for looking beyond the design of conventional menu-based dialogues to explore alternative and more flexible means of offering users access to services. For example, an enterprise may want to introduce new informational or transactional services that may not be considered in the initial application design under normal circumstances, such as short-term offers or product promotions, but at the same time avoid adding these as options to menus which may become unnecessarily long and complex.

One solution for adding new options in a menu-driven service is to introduce a short system-initiated informational message within the dialogue structure with the intention of disseminating new information relevant to a particular customer at a specific point in time during their use of the automated service. This system-initiated message could simply consist of a brief prompt which may or may not be followed by a short dialogue (e.g. requesting a yes/no response) enabling the user to pursue or decline the offer immediately. The system-initiated message interrupts the regular turn-taking of the dialogue and, in doing so, impedes the human user from continuing with the flow of the call as anticipated. These messages may therefore be viewed as a particularly pronounced form of System-Initiated Digressions (SIDs) since they are in effect unsolicited, unexpected and not directly related to the current topic or the prime goal of the call.

Mass-market automated services are primarily designed to handle task-driven conversations within a narrow topic domain, such as flight information, banking account transactions or cinema bookings. The user of such services typically expects the interaction to be restricted to the chosen topic and task at hand and that the computer will co-operate fully to complete the goal of the call. Fixed turn-taking, goal-driven, prompt-response interaction has become the conventional way of designing automated self-service telephone applications. It is not common practice for an automated service to initiate an interruption or launch into new topics, a fact which may explain why such dialogue behaviour remains largely unexplored in the current literature for spoken human-computer interaction. The possibility of deploying system-initiated digressions in human-computer conversation raises new and interesting dialogue engineering issues regarding the design, usability and acceptability of such applications. Successful strategies in this area could have important positive commercial implications.

The research described here explores how a SID may be used to disseminate unsolicited financial information – or offers – in a speech-driven automated telephone banking service, along with hints to the user about how to pursue the offer. These digressions work by interrupting the user and suspending the regular dialogue turn-taking for the duration of the informational message or proposal. The SDS under investigation is an automated telephone banking application, however, the issues raised in this research are relevant to the implementation and development of most similar mass-market self-service SDSs which rely on menu selection.

1.4 Thesis outline

The aim of this thesis is, through a body of empirical research, to define strategies for devising and deploying system-initiated digressive dialogues in an already deployed mass-market self-service dialogue system. As stated in the introduction, such dialogue behaviour has not yet been fully addressed in the current literature; the purpose of the literature review in Chapter 2 is therefore to identify related areas of research and to explore how current findings and theories may impact the design of system-initiated digressions in auditory-only interfaces. The approach employed in this research is multidisciplinary, bringing together research areas such as dialogue engineering, anthropomorphic human-computer interface design, linguistic theory and usability evaluation.

Crucial to the current research is to obtain feedback from potential users who will experience these digressions while using the self-service application to carry out some banking tasks. A series of four experiments were devised in order to capture users' impressions of the digressions and to investigate how their perception of the service usability may be affected by this novel dialogue behaviour. Chapter 3 provides details of underlying dialogue-engineering principles which motivated the research and also gives an overview of the experiment setup and methodology employed. The four ensuing chapters give further design details for each of the experiments and are centred around four main dialogue engineering themes for the deployment of system-initiated digressions: 1) the turn-taking strategy employed, 2) the location of the proposal, 3) the register used in the interruption and 4) cognitive aspects involved in introducing the concept of 'hidden' menu options.

Chapter 8 gives a summary of the main findings from each of the four experiments and provides a discussion of the implications of the findings in the field of dialogue engineering. Finally, further areas of related research are identified and topics for additional experiment studies are proposed.

Chapter 2

You've got to know where the machinery is and how it works before you can throw a monkey-wrench into it.

- Michael H. Brown, in 'Brown's Lawsuit Cookbook' (1981) -

Literature review

2.1 Introduction

The advent of speaker-independent speech recognition can be held to have added a new dimension to mass-market automated telephone services, offering the potential of increased usability for callers. Speech input presents a number of advantages over the traditional touch-button entry method (Rosenfeld et al. 2001; Halstead-Nussloch 1989; Whittaker & Attwater 1996). Firstly, the placement and size of the keypad on the phone handset often make it more convenient, and less distracting, to speak responses rather than pressing buttons. Secondly, particularly for inputs such as airports or cities, it is usually more suitable to refer to objects and services by their actual names rather than mapping these onto an arbitrary touch-button combination. Thirdly, speech recognisers allow for the use of synonymous expressions, such as (in response "When would you like to travel?") "tomorrow", "on Saturday", "in three days time", or "on the 15th of June". A change from touch-tone to speech operated telephone services may have important commercial benefits; Suhm et al. (2002) compared touch-tone to natural language input in a telephone call routing[1] system and found that customers preferred the speech input version.

Dialogue engineering is a relatively young subdiscipline within the field of human-computer interaction, a field which has been heavily dominated by research and development of screen-based Graphical User Interfaces (GUIs). In fact, most of the standard textbooks on interface strategies and guidelines only give a short account of the application of speech recognition technology in computer interfaces (Preece et al. 1994; Dix et al. 2004; Shneiderman 1998), or include just a brief mention of the topic (Nielsen 1993; Preece et al. 2002; Raskin 2000). Even though some of the guidelines and human factors presented in the GUI-related literature also apply to the design of Voice User Interfaces (VUIs) they are often of limited applicability due to the

[1] Call routing involves associating a request (e.g. "I want to buy a house") with a desired destination (e.g. mortgage/lending department in a banking service).

fundamental differences presented by the visual and auditory media channels. The visual, screen-based interface offers the ability to present multiple pieces of information, permanently, simultaneously and to give them contrasting prominence (e.g. flashing graphics, bold colours). The auditory-only interface is serial, transient and paced which limits the way information can usefully be conveyed to the user (Yankelovich 1995; Dybkjær & Bernsen 2000). In GUI, the case for speech has been less convincing as the user is seated in front of a computer equipped with keyboard and display (Lai 2000); auditory input/output has proven to be of limited use and is mainly employed under specific circumstances (e.g. speech recognition in hands-busy situations or as a tool for individuals with special needs such as dyslexia, visual impairment or for research purposes).

Little has been published in terms of practical guidelines on how to develop VUIs for dialogue engineers – that is, until the past 5-10 years or so. It now appears that the concept of the 'speech interface' – with particular emphasis on dialogues for mass-market telephone services – is now sufficiently ubiquitous and commercially applicable to have earned VUI status as a research field in itself. Several books have been published recently on this topic, offering practical guidelines and providing broad coverage on how to implement such services (Cohen et al. 2004; Gardner-Bonneau 1999; Kotelly 2003; McTear 2004; Weinschenk & Barker 2000). However, they offer little in terms of the maintenance and future development of such services once they have been deployed; nor do they address issues of introducing new or less frequently requested service options. Marics & Engelbeck (1997:1099-1100) briefly address the issue of adding/removing items in touch-button operated services, stressing the importance of accounting for experienced 'power' users (who may interrupt prompts and miss out on changes) and of preserving the current mappings between action and assigned key-press for existing service options.

2.1.1 Chapter outline

As already stated in the previous chapter, this research will explore the use (design and consequences) of System-Initiated Digressions (SIDs) as a means of introducing new service options into the dialogue of an already existing automated telephone banking application. This is a novel dialogue behaviour which has not been fully explored in the research literature to date; thus, it becomes the purpose of the current research to fill this gap in knowledge. The approach taken is threefold. Firstly, to define the current state-of-the art in commercially available speech technology and pertinent mass-market dialogue engineering techniques. Secondly, to identify related disciplines and explore their relevance to the topic of SIDs. Thirdly, to design, implement and empirically evaluate SID dialogue strategies by means of four large-scale usability experiments.

The remainder of this chapter begins with an overview of the components of a typical spoken language system devised for building automated telephony applications. A review of current practices in dialogue engineering for the design of system prompts is then given, followed by a closer examination of how digressions and interruptions are currently implemented in human-computer interaction. Politeness theory, intimately linked to the issue of interruptions in dialogues, is introduced and the related controversies associated with endowing the computer interface with such predominantly human characteristics (anthropomorphism) are explored. Finally, an account of current practices in the design and evaluation of usability in spoken dialogue systems is given.

2.2 Components of a spoken dialogue system

Mass-market spoken dialogue systems perform, in general terms, three main actions: they generate output (prompts and messages), recognise speech input from the caller and perform actions with the aim to move the conversation towards the caller's goal. The end result is a (hopefully) successful conversation between a human and a machine where relevant information has been obtained or appropriate actions have been performed. There are many different architectures used for dialogue systems. Figure 1 shows the components of a typical spoken language dialogue system (as employed in the thesis) featuring a fixed, non-adaptive scheme. The actual number of components, their associated responsibilities in the recognition process and how they are linked together differ from system to system.

2.2.1 Pre-processing and feature extraction

In the **pre-processing** stage, the system performs echo cancellation in order to filter out any echo introduced by the telephone line. This improves the quality of the speech signal and, in the case of barge-in, is necessary to prevent the system from mistaking its own prompts for caller speech input. Secondly, the pre-processing detects the start and end of the caller's speech input (dealing with leading or trailing silences or background noise); the resulting utterance waveform forms the input to the **feature extraction**. The feature extraction process in Nuance (the commercially available speech recognition system used in the current research) allows a certain amount of background noise to be removed from the waveform and typically segments the audio data into frames (e.g. 10-milliseconds) for analysis. The resulting *feature set* from this process consists of feature vectors (a list of numbers representing measurable characteristics in the speech) which in turn provides the input to the **recognition search**.

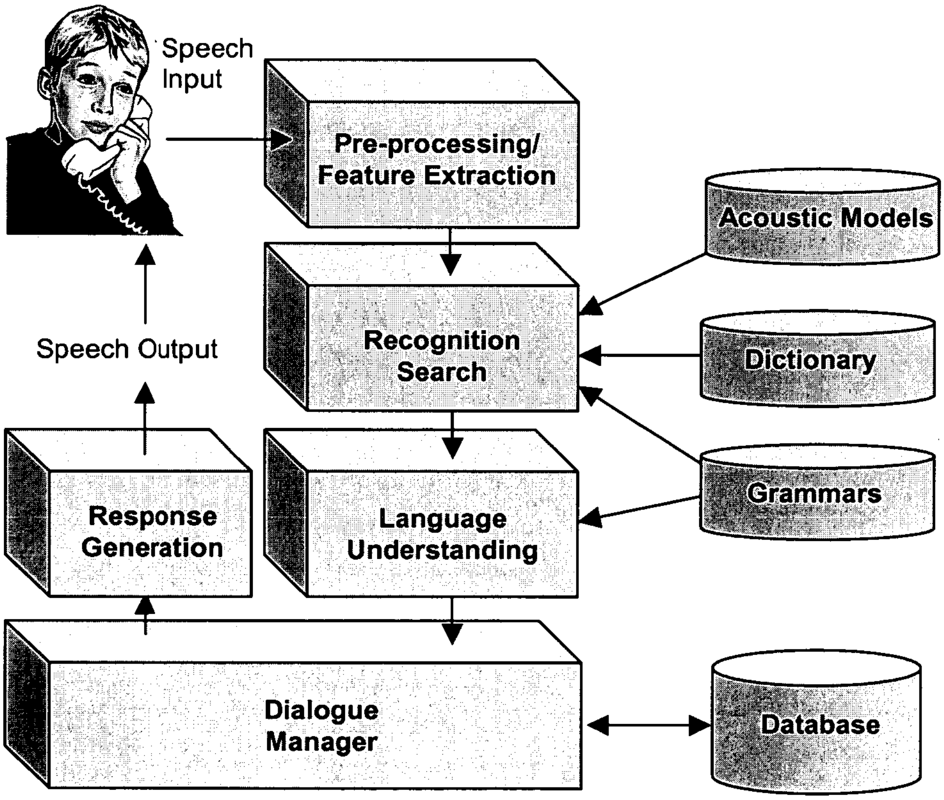


Figure 1. The main components of a typical fixed, non-adaptive spoken language dialogue system as featured in the experiments Chapters 4-7.

2.2.2 Recognition model and language understanding

The next step is for the **speech recogniser** to interpret the feature vectors and produce a transcription (text) of the caller's speech. At the lowest level, the **acoustic models** holds representations of all possible phonemes in the language. A phoneme is the smallest unit in a language that is capable of conveying a distinct meaning such as the *m* in *mat* and *c* in *cat* in English. Phonemes can be context dependent in that their pronunciation may be affected by surrounding phonemes (coarticulation) which is reflected in the acoustic models. Each language handled by the recogniser requires associated acoustic models which are developed by training the system with multiple examples of how phonemes in context are pronounced.

Phonemes, in turn, are concatenated to make up words. The **dictionary** consists of a list of word entries paired with their associated pronunciations (phoneme sequences). Some words have multiple entries in order to account for different possible pronunciations and regional variation.

Finally, the **grammars** combine the words in the dictionary into phrases to define the entire set of word strings that the system can recognise. Different grammars are used at different stages of the interaction. For example, a grammar containing bank account names is used for the system prompt "Which account are you interested in?", and a grammar defining pounds and pence input is used for "What amount are you looking for?". Grammars can also specify extraneous information in the caller's speech such as filler words and phrases: "<u>Uhm</u>, <u>I'd like to</u> transfer some money, <u>please</u>." Grammars are usually rule-based (finite state network) or founded on a Statistical Language Model (SLM). The rule-based (deterministic) grammar is crafted by writing explicit representation of phrases that define the grammar. The SLM (probabilistic) is developed by feeding a large set of transcribed utterances to a system which produces a grammar by calculating the probability of a particular word occurring in a particular context. The SLM generally allows callers greater flexibility in what they can say, but requires an extensive collection of transcribed utterances on which the model can be based (Furman 1999).

The combination of a grammar, associated dictionary and acoustic model forms the speech recognition model with representations of all possible word strings and all their possible pronunciations. The feature vectors, which represent the caller's speech, are then compared to these representations in order to find a best match; the result (word string) is what has been recognised. Often the recognition system returns a 'confidence measure' along with the recognition result (or more than one result in an N-best list of possible matches) to define how closely the caller's utterance corresponded to the representation held in the recognition model.

In the **language understanding** process, meaning is assigned to the caller utterance. A common way of representing the meaning of an utterance is by using slots (or place holders) in the grammars which may then be assigned a specific value in the recognition process. For example, the string representation of a possible input (with slots) may look like this: "I want to transfer <amount> from <source account> to <destination account> on <date>". The slot values form the meaning of the caller's input and the dialogue manager acts upon this information.

2.2.3 Dialogue manager

The **dialogue manager** consists of a program written to control the flow of the interaction. In current commercial systems, the implementation of the dialogue manager often involves employing tools and application programming interfaces provided by the platform vendors, or by using purpose-defined languages such as VoiceXML[2]. The dialogue manager coordinates the modules in the system, specifies when the system should start listening for input from the user, and what to listen for. The dialogue manager takes action based on the recognition result: accesses external information such as **database** contents (e.g. customer account information), plays information to caller (e.g. the account balance), performs transactions (e.g. transfers money) or prompts for further details from the caller (e.g. "Would you like to arrange another funds transfer?").

[2] Further information can be found on the VoiceXML Forum website: http://www.voicexml.org/.

2.2.4 Response generation

The system output is generated by playing recordings of human speech or by employing text-to-speech (TTS) conversion. Although TTS technology has improved in recent years, the TTS output does not yet match the quality and intelligibility of recorded human speech. In general, recordings of human speech are mainly used in commercial applications augmented by TTS technology when required, particularly for dynamic text data such as e-mails and news readings. The design of system prompts is usually considered more an art than science, relying on expert intuition and tacit experience (Hansen et al. 1996).

2.2.5 Further speech technology features

There are additional available features in commercial speech recognition systems that may be added to the component overview described above. Firstly, the use of barge-in (or cut-through) technology enables the caller to interrupt system prompts and can therefore provide a faster interaction. Secondly, speaker verification technology can be used to verify that a caller is the person he or she claims to be. When speaker verification is used, the caller needs to complete an enrolment process in which voice data are stored in a model of the person's voice. This model – or voiceprint – can then be used in future calls to verify the callers claimed identity.

2.3 Dialogue engineering

Whilst the capability and accuracy of speech recognition technology are continuously improving, the need still exists for the system messages (or prompts) to be designed to constrain the range of user inputs to those that match the capabilities of the speech recognition grammar (Bernsen et al. 1996, Karis & Dobroth 1991, Tomko & Rosenfeld 2004). It is the task of the dialogue engineer to design the system output (prompts) and to ensure that the recognition engine can successfully process a suitable range of user responses.

Dialogue engineering for speech driven mass-market applications is mainly concerned with development issues relating to the technology at hand, such as whether to use voice recordings or text-to-speech for system prompts; whether to allow callers to barge-in during system prompts; whether to use open or closed prompt styles (Hone & Baber 1999); whether to use isolated word recognition or allow for more fluent speech; and whether to allow universal commands such as "cancel" or "exit". These are all pertinent issues that need to be resolved at the early stages in the application design; primarily depending on the recognition technology used and criteria specified in the requirements capture, such as the skills of target users, the service domain of the application or type of output.

Above all, special care needs to be addressed to the design of the system's informational messages and prompts; in a telephony application, this is the only mode available through which to convey to the caller what to say, how to say it and what the options are. Thus, as far as the user is concerned, the system output *becomes* the interface. One of the biggest challenges in creating a spoken automated dialogue system is to convey to the user the system capabilities, the range of allowable speech inputs and the domain knowledge (Glass 1999). The aim is to create a 'habitable' interface in which users are allowed to express themselves adequately (Hone & Baber 2001; Green 2002).

2.3.1 Choosing the voice talent

The choice of speaker for the system prompt requires careful consideration (ETSI[3] ETR 329). The most important consideration is how the voice (loudness, intonation, speed and rhythm) comes across over the telephone channel. To achieve this, professional speakers are generally used and the recording sessions are often carefully coached. Selecting the appropriate voice is not only important from the user's cognitive and interpretative skills point of view; the system voice also reflects the 'company image' and may have a significant impact on customer attitudes towards the service.

[3] ETSI (European Telecommunications Standards Institute), see website http:V/www.etsi.org/.

2.3.2 Prompts and menus in mass-market SDSs

The system functionality remains hidden to the user in auditory-only interfaces; a central concern in the design of such applications is therefore how to let callers know about the range of available options that they may choose from (Yankelovich 1996; Kamm et al. 1998). Several methods and prompting styles have been proposed. Typically the design efforts have centred around three main strategies: 1) 'open-ended' prompts (e.g. "How may I help you?") inviting the user to say any utterance; 2) 'closed-set' prompts where the user makes a selection from an up-front list of options; or 3) 'on-request' prompts where the user is not presented with options unless specifically requesting this information, for example by using commands such as "help" or "hear list".

Several studies have been conducted to explore how contrasting menu prompt strategies may affect user satisfaction and task completion in applications such as: call routing (Sheeder and Balogh 2003, Williams et al. 2003a, 2003b; Witt & Williams 2003; Williams & Witt 2004), telephone directory assistance (Vanhoucke et al. 2001), telephone banking (McInnes et al. 1999), e-mail (Walker et al. 1998) and newspaper subscriptions (Dialogues 2000 Report, 1997).

There is no universally applicable strategy to draw on when designing prompts for voice-driven telephone applications. The ideal strategy will depend on the skills of the intended user group, the technology capabilities at hand, the application domain, the frequency of use and the complexity of the underlying data structure (Vanhoucke et al. 2001). Trade-offs are associated with each strategy: the open prompt strategy is short but may cause users to be confused or unsure of what to say; the closed menu prompt guides users to what they can say and how to say it, but menus can be long or involve a complex hierarchical navigation structure. In the menu-driven approach, users may sometimes also find it difficult to match their goals to the options presented to them.

There are several examples of research into prompt strategies in the current literature. Open-ended prompt strategies are often employed in call routing applications and usually involve obtaining a large corpus with transcriptions of hand-routed calls to train the recognition engine by means of statistical language modelling (see Section 2.2.2 above). Carpenter & Chu-Carroll (1998) and Lee et al. (2000) employed an open-ended "how may I direct your call?" for incoming calls to a financial institute. After training and developing their model for routing the calls they found their system to perform roughly at the same level of accuracy as human operators. Gorin et al. (1997) provide another example of routing responses to "how may I help you?" in a different call router domain (to handle enquiries to a telecommunications provider).

A problem with an open-ended phrase is that callers may remain silent or hesitate, particularly if they are aware that they are communicating with a computer and are unsure what it 'can understand'. An open general prompt such as "How may I help you?" in a travel application is likely to elicit overly general and uninformative responses, such as "I want to plan a trip" (Stallard 2001). Other work (Walker et al. 1998) compared two computer prompting styles in an e-mail telephone application: mixed-initiative ("Hi Elvis here. I've got your mail.") and system driven ("Hi, Elvis here. You have 5 new and 0 unread messages in your inbox. Say read, or summarize, or say help for more options."). Results showed that although the mixed-initiative strategy was more efficient (number of turns, task completion times) users preferred the system-initiative interface: the additional flexibility of the mixed-initiative interface leads to user confusion about their available options and poor performance by the speech recogniser (Walker et al. 1998). In contrast to these findings, research (Hone & Baber 1999) to compare the use of open-style ("please state the service you require") and menu-based prompts ("which service do you require, balance enquiry, cash transfer or other service?") arrived at the conclusion that performance gains which can be expected from imposing high levels of dialogue constraint (increased recognition accuracy) are relatively small compared to the costs of imposing such constraint (longer transaction times).

To constrain the user input when an open prompt is used, one solution is to add input hints. For example, the prompt "How may I help you?" can be made more specific by suggesting to the user the type of utterances which are expected "Which *service* do you require?" (see McInnes et al. 1999, Cohen et al. 2004). To support the caller further, the prompt may include examples of keywords or phrases that the system can recognise; work in developing a call routing system for a (fictional) wireless telephony carrier (Sheeder & Balogh 2003) revealed that correct call routing was significantly higher when callers are exposed to an initial prompt with an example phrase which preceded the initial query. It was also found that, as a consequence of using keyword examples, some participants tended to repeat back a keyword, even when this was not a correct representation of the task at hand. Tomko & Rosenfeld (2004) explored the use of "speak simply" instructions (deploying short, medium and long versions) to convey to the caller the recognition capabilities of a telephone service for cinema and flight time information. They found that the length of the instruction had an impact on number of words used in each utterance: as the instruction become longer and more specific the participants tended to use fewer words in their requests.

The majority of commercially deployed mass-market automated applications are aimed at servicing the general public and must therefore provide for callers with diverse levels of experience and knowledge. In these types of self-service applications, the use of closed-set prompts in the form of explicit menu listings is the most popular and most frequently employed method for informing users (especially novice users) about the range of services available to them; it has become the *de facto* standard (Resnick & Virzi 1992). Although listening generally requires less perceptual and cognitive effort than reading (Preece et al. 2002), information presented through the auditory-only interface is serial, transient and paced, which puts a strain on users' cognitive and perceptual resources. This has an impact on the length of system prompts and limits the number of options that can usefully be presented in menus[4]. Another issue with menu-driven services is that power users may get fed up with having to go through the list of options each time they call the service. One solution is to allow for barge-in of prompts. A further option is to combine open-ended and closed-set strategies, such as presenting the caller with an

[4] Referred to as the 'echoic load' and is based on limits in human working memory; a maximum of 3-4 elements are usually recommended for voice menus (ETSI 202 116). For further information on cognitive load design consideration for voice interfaces, see Cohen et al. 2004: 119-131.

open prompt initially and then, if the speech recogniser detects no response (silence), playing the list of menu options or giving further instructions (Sheeder & Balogh 2003).

Finally, particularly pronounced forms of 'closed' prompts are employed in dialogues where user inputs are restricted to responding with "yes" or "no" to each menu option or, as in the 'skip and scan' (or sometimes called 'zap and zoom' strategy), by using a predefined command, e.g. "next", "previous" or "select" (see Hornstein 1994, Dialogues 2000 Report 1997, Goldstein et al. 1999; Krüger & Kruckenberg 1999; Resnick & Virzi 1992). One problem with this strategy is that users first need to get familiar with the commands which drive the application. One possible solution is to include a message at the start of the call which explains the use of the skip and scan navigation (Hornstein 1994).

2.4 Digressions in dialogues

In the current literature, there are virtually no investigations of the introduction of digressive dialogues into human-computer interaction. In the broader context of human-human conversation, digression is the term used to describe a part of a discourse in which a interlocutor introduces information that is not directly related to the current topic or main goal of the conversation. Such digressions are common in human-human conversation where participants use their knowledge about coherence, states of attention and intention in the discourse to find the appropriate timing to introduce new topics into the flow of a conversation (Lenk 1998). This intrinsic human ability to co-ordinate and collaborate in interactive activities poses a challenge to designers of human-computer interfaces, particularly in the area of mixed-initiative interaction (Haller & McRoy 1997; Horvitz 1999).

The limited research that has been conducted into digressive dialogues in spoken human-computer interactions (often referred to as 'out-of-turn interaction' or 'unsolicited reporting', Hearst et al. 1999) has focused mainly on providing models for handling *user*-initiated digression (Ramakrishnan et al. 2002; Narayanan et al. 2000; Haller 1994) which occur when the user supplies extra or out-of-turn information in response to system prompts. This new or extra information supplied by the user is however normally related to the overall goal of their participation in the conversation.

Digression, in the form of informal 'small talk', has been investigated in human-computer interaction with the aim of building rapport with the user. For example, a virtual agent – called 'Rea' – used in a speech-enabled graphical property purchasing application, initiated small talk about the weather with the aim of demonstrating the expertise of the agent and to directly or indirectly satisfy task goals (Cassell & Bickmore 2000; 2003). Results from a pilot study indicated that users trusted and liked 'small talk Rea' more than the version when no small talk was generated. System-initiated digressions have also been used to introduce additional (but relevant) details in response to user-prompted yes/no questions in human-computer dialogues (Green & Carberry 1999). This model was tested in an experiment where a user who is outside and cannot see inside the laboratory communicates with a robot which is inside the laboratory. The users had to request information from the robot by means of a questionnaire with 11 yes/no questions. They found that responses that included extra information to the answer (rather than just yes/no) resulted in enhanced user attitude towards the service.

The potential for introducing system-initiated digression into dialogues for automated telephone services – the domain of this research – remains largely unexplored.

2.4.1 Interruptions

The dialogue in mass-market automated telephone services is typically scripted and does not change between phone calls to the service. Therefore, a digression in such a context not only introduces a new topic into a dialogue – it constitutes an interruption of the dialogue which suspends the scripted turn-taking for as long as it takes to deliver the message information. Such interruptions can be both distracting and disruptive to the user, particularly if the digressive information is unsolicited and unwanted as witnessed by the discussion about the controversies associated with the hints and tips offered by the Microsoft 'paperclip' (in MSN BC 1999[5]).

McFarlane (1997, 1998) provides a comprehensive overview of the issues concerned with interrupting human-computer dialogues. The interruptions are mainly addressed to the multitasking graphical environment, however he identifies a taxonomy of 8 factors for coordinating interruptions which have general applicability: 1) source of interruption, 2) individual characteristics of person receiving interruption, 3) method of coordination, 4) method of interruption, 5) method of expression, 6) channel of conveyance, 7) human activity changed by interruption, 8) effect of interruption (McFarlane & Latorella 2002). The concept of 'coordination' in this context implies that any interrupting event (such as a warning or an alert) needs to be harmonised with competing processes (such as user's current task or his/her cognitive ability to multitask). For example, some interrupting warnings may take precedence over all other current user activities while other less critical interruptions may be deferred until the user has completed the primary task.

Bailey et al. (2001) explored the impact of peripheral interruptions (non-essential information that may be of interest or helpful to the user but not necessarily related to the user's current task) in a graphical task environment for adding, counting, image comprehension and reading comprehension. They found that users performed slower, experienced greater anxiety and perceived the primary tasks to be more difficult when they were interrupted. The level of annoyance depended on the category of the primary task and the timing of the peripheral task. Similarly, Cutrell et al. (2001) studied the influence of instant messaging in a book search graphical tool and found that these types of interruptions can appear "...disruptive, both frustrating users and decreasing the efficiency with which they perform ongoing tasks". A related study has shown that the disruptiveness of instant messaging can be reduced if the incoming message is highly relevant to the current task, or if the message is queued until the primary task has been completed (Czerwinski et al. 2000). Oulasvirta &

[5] Available at: http://zdnet.com.com/2100-1107-513612.html?legacy=zdnn (link last checked 18/06/05)

Saarilouma (2004) devised an experiment to explore the impact of interrupting messages on listeners' long-term working memory; the interruptions occurred while participants listened to stories presented over a video channel. Results indicated that semantically similar interrupting messages can disrupt listeners' ability to remember the original text, more so than when the interrupting message is from another domain.

Interruptions in the graphical on-screen environment are often facilitated by the fact that the user can be supported by visual cues regarding the task that they were engaged in prior to the interruption. Further, the screen can be used to present information simultaneously and give users cues of different strengths regarding the importance of the incoming interruption. The auditory-only interface, however, has no left-to-right, no ability to 'click' since 'here' passes by with the uttered phrase; the ear cannot 'browse' and becomes confused when presented with simultaneous info – the 'cocktail party effect' (Arons 1991). Interruptions in auditory-only interfaces are therefore likely to have a greater impact on the user's attentional cognitive resources, increasing the disruptiveness and the probability of mental mistakes. To support efficient task recovery, the process of interrupting can be divided into three phases (Franke et al. 2002): 1) pre-interruption: switch to a different voice; 2) mid-interruption: interrupt if critical, alert if equally important to current task, hold off alert if less important to current task; 3) post-interruption: provide recovery support by supplying commands or allowing user to request a summary of actions performed. The strategy has been tested in a voice application for supporting Marines in managing their requests for supplies (Franke et al. 2002, field tests are currently pending).

2.5 Politeness Theory

2.5.1 Politeness in human-human interaction: Brown and Levinson's theory of politeness universals in language usage

Politeness in human communication has received much attention in the field of pragmatics and sociolinguistics over the past two decades and has mainly been focussed on how communicative strategies are employed in order to promote and maintain social harmony[6] in human-human interaction. One of the most influential and best known theories of politeness is that developed by Brown & Levinson (1987). Their politeness theory is based on the notion that each individual has positive and negative 'face wants' and that these are ascribed by all (rational) interlocutors to themselves and to one another in any social interactive situation. Two face wants are defined:

> NEGATIVE FACE: the desire to be un-impeded in one's actions, the basic claim to territories, personal preserves, rights to non-distraction – i.e. freedom of action and freedom from imposition.
>
> POSITIVE FACE: the desire (in some respects) to be approved of, the positive consistent self-image or 'personality' claimed by interactants (crucially including the desire that this self-image be appreciated and approved of).

Any utterance or action in a communicative situation can be seen as potentially threatening to the positive or negative face of either of the interlocutors and, consequently, expressions of politeness are normally used as mitigations aimed at redressing this threat. Although Brown & Levinson give examples of how the speaker's *own* positive and negative face wants may be at risk[7], much of their politeness theory is primarily focussed on the explicit strategies used by a speaker to avoid damaging the addressee's face wants (Chen 2001). The speaker uses politeness expressions to indicate that no face threat is intended or desired, and to convey that the addressee's face wants are recognised and approved of by the speaker.

The relative 'seriousness' of a face-threatening act are based on three 'social dimensions' (Brown & Levinson 1987:74). These are: the relative power of the addressee over the speaker; the social distance between the speaker and the addressee; and the ranking of the imposition involved in doing the face-threatening

[6] For an account of impoliteness strategies see Culpeper (1996).

[7] For example, expressing thanks or making an excuse are damaging to the speaker's negative face. Admissions of guilt or non-control of emotions (laughter or tears) are examples of damage to a speaker's positive face.

act. Each of these dimensions is context-sensitive, meaning that the relationship between two individuals (such as the relative power of a manager over an employee) may be inverted under certain circumstances. Depending on the seriousness and social setting for the face-threatening act, a number of options are presented to the speaker on how to redress a potential face threat. First of all, the speaker has the option of not performing the act at all and could therefore theoretically avoid damaging the face of the addressee altogether. However, if the speaker decides to go ahead with the face-threatening act, Brown & Levinson identify a taxonomy of politeness which includes four principal categories of expression strategies: (1) doing the act without redressive action (baldly), (2) using positive face-redress, (3) using negative face-redress, or (4) doing the act off-record. The 'off-record' strategy attempts to minimise the face threat by creating uncertainty as to the existence of the face-threatening act itself, for example, by using ambiguous or vague expressions, or by using hints such as "it's cold in here" (implying "shut the window").

To carry out an act 'baldly', without redress, involves doing it in the most direct, clear, unambiguous and concise way possible. The speaker may use the bald strategy when there is no fear of retribution by the addressee (for example in the interest of urgency or efficiency, e.g. "watch out!"); where the danger to the addressee's face is very small (such as in proposals and requests); or where the speaker is considerably superior in power to the addressee.

Face-redressive politeness strategies are used when there is a perceived potential threat in an utterance to either the positive or negative (or both) face wants of the addressee. Utterances that are considered threatening to the *negative* face wants of the addressee will include: ordering the addressee to do something, making an offer which may incur debt for the addressee or expressions of strong emotions towards the addressee. Negative face-redressive strategies are characterised by formality and distancing. It is such forms of 'negative politeness' that are conventionally associated with politeness in everyday language, such as "excuse me" and "thank you", as these relate to the imposition itself. *Positive* face-redress, on the other hand, widens the sphere of politeness to include the appreciation of the addressee's wants in general or to the expression of similarity between speaker's and addressee's wants. Threats to the addressee's positive face wants are caused by, for example, bringing bad news about the addressee, expressing disapproval or raising emotionally divisive topics. The positive face-redress strategy is characterised by 'intimate' language behaviour and makes reference to a close interdependent social relationship between the interlocutors. For example, the speaker might use in-group identity markers ("hey buddy") or show intensified interest in the addressee's wants ("your hair looks *great*").

Some face-threatening acts, such as interruptions[8], are considered to be intrinsically threatening to *both* the negative and positive face wants of the addressee (Brown & Levinson 1987:67). An interruption constitutes a threat to the negative face wants of the addressee because it infringes to some degree on the addressee's right to non-distraction and desire to be un-impeded in their actions. Interruptions also pose a threat to the addressee's positive face wants by implying that the person who interrupts ignores or does not care about the addressee's feelings and wants.

2.5.2 Some critique of Brown and Levinson's theory

It is worth noting, at this point, that Brown and Levinson's theory is not the only prevailing attempt at identifying and defining the politeness phenomenon[9] and there is no real consensus in the current literature regarding a formalised concept of politeness. What can be said is that their theory is one of the most established, comprehensive and well-referenced theories on the motivation for using politeness registers in spoken utterance to date. Over the years, however, their politeness theory claims have received some criticism regarding the validity and applicability. This section will address some of these issues but by no means represents a full account of the wealth of literature published on the topic.

[8] Other face-threatening acts considered to intrinsically threaten both the negative and positive face wants of the addressee are complaints, threats, strong expressions of emotions and requests for personal information.

[9] Watts (2003) provides a comprehensive review of some of the issues surrounding politeness in linguistics with details of competing strategies.

One of the main critiques raised over Brown & Levinson's theory is that it takes an anglo-centric approach to face and is not proven to be applicable to investigations carried out on different cultures (Escandell-Vidal 1996; Meier 1995b). It has also been noted that their politeness theory is too focussed on the individual's face threat (Spencer-Oatey 2002); in some cultures it is the individual's status within the group that takes precedence over the actual individual's own needs/wants. This would make their taxonomy inadequate for cross-cultural analyses. However, other politeness researchers have argued that the concept of positive and negative face is indeed universal – it is the relative *importance* of positive and negative face that differs between cultures (O'Driscoll 1996).

Another issue centres on claims that Brown & Levinson's theory is too focussed on the face wants of the listener (or 'other-face') and does not take into account the 'self-face' of the speaker (Chen 2001; Meier 1995a). Chen (2001) gives the example of the act of stepping on someone's toe and claims that (according to Brown & Levinson's taxonomy) the ensuing act of apologising is an act that threatens the speaker's self-face (i.e. the expression of regret, admission of guilt). According to Chen, the most 'polite thing' in this circumstance would (following Brown & Levinson's taxonomy) be not to perform the ensuing self-threatening act of apologising. Contrary to Chen's line of reasoning, it could be argued that the face-threatening act is actually caused by the imposition to the other person's negative face, caused by stepping on the toe (albeit accidental). The most polite thing would have been to avoid this social faux pas altogether and thus avoid the face-threat, or – now that it has already happened – duly apologise. Despite having issues with the definition of self-face threats, Chen (2001) concludes that Brown & Levinson's theories are "fundamentally correct and is still the best tool ... in the investigation of politeness".

2.5.3 Politeness in human-computer interaction: issues of anthropomorphism

Politeness is undoubtedly an important aspect of *human-human* conversation, but little prior work has been undertaken to investigate how relevant it is to *human-computer* dialogues. What are the conversational rules or social dimensions that govern the use of politeness registers in dialogues where one of the interactants is a computer? Can existing politeness theories be expanded to encompass human-computer interaction? If so, what politeness strategies should the computer (in the capacity of the speaker) be endowed with, and how are the resulting politeness expressions received by the human user?

People's interactions with computers (and other media) are fundamentally social (Reeves & Nass 1996; Nass & Moon 2000; Nass et al. 1994). This view is founded on the notion that the human brain has evolved to respond and relate socially to human-like entities in our surroundings and that this innate reaction is almost impossible to overcome (without conscious effort) – even in situations where humans interact with a supposedly non-social entity such as the computer. This propensity for humans to relate socially to media has been explored in a series of controlled experiments (Reeves & Nass 1996; Nass & Moon 2000; Nass & Lee 2001). The results showed that users applied gender stereotypes to computers; they identified with computer agents sharing their ethnicity;, imputed personality to computerised voices; and they were more attracted to agent characteristics (submissive/dominant) that were similar to their own personality. Experiment studies have also shown that participants are sensitive to consistencies in computer character personality (introvert vs extrovert): consistency between computer voice and text in a book-shopping website was desirable in order to maximise social presence in media (Lee & Nass 2003); and participants preferred computer characters with consistent verbal (text) and non-verbal (posture) cues (Isbister & Nass 2000).

The experiment results obtained by Nass and colleagues strongly suggest that human users have a subconscious tendency to apply deeply-rooted social rules to interactions with computers in the same way as they do when interacting with other fellow humans. These social rules seem to relate to our innate disposition and cultural upbringing. But how do users react to a computer that blatantly attempts to exploit these social rules? Fogg & Nass (1997) explored the effects of employing computer-initiated flattery when giving feedback to users in a text-based guessing game application. Experiment results showed that flattering feedback (compared to the generic feedback condition) had a positive effect on a number of aspects of the interaction. For example, the flattery increased participants' feelings of power; made them more positive towards their own and the computer's performance; and made them enjoy the interaction more.

Nass and colleagues concluded that users apply 'over-learned' social rules to computers, such as politeness: experiment results showed that participants gave a significantly more positive evaluation of a computer's performance when questioned directly by the computer itself compared to when questioned by a different computer or through pen and paper questionnaires. This would indicate that politeness is an important factor in human-computer interaction. However, the work on politeness in human-computer interaction has been centred around how humans behave politely towards computers, rather than investigating how humans respond to a computer that tries to adopt and communicate polite behaviour more explicitly.

The use of politeness in system-initiated interruptions has been explored in the context of a graphical library search engine interface (Colón et al. 2001). These interruptions involved on-screen error text messages (resulting from either system errors or user errors) that were presented with or without politeness (courtesy). The messages were deployed in the library application and evaluated in a controlled experiment. The two main findings from the experiment were: firstly, the interruption performed by the computer interface had a detrimental effect upon the user perception of the interaction with the computer (the participants judged the interaction as being less friendly, less motivating and less beneficial). Secondly, it was found that politeness strategies had no effect on minimising participants' negative reaction towards the interruption. Similar results have been obtained (Tzeng 2004) in research where apologetic vs non-apologetic feedback and emoticons (sometimes referred to as 'smilies') where used in a graphical computer guessing game with text input/output; the apologies and emoticons neither helped to improve the subjects' guessing performance nor did they prevent the participants from feeling bad about their performances – but they made the program more aesthetically desirable.

Ribeiro & Benest (2002) explored the use of voice interruptions in the form of spoken notifications and reminders (from email, printer and diary agents) in a graphical interface environment. The interruptions applied a certain level of anthropomorphism by using linguistic variation and expressions of politeness. 15 participants took part in the experiment and were told to use the World Wide Web to find answers to 48 questions. Results showed that participants agreed that the interrupting messages were polite and slightly agreed that the polite messages were appropriate. Polite attention-getters were received with mixed emotions; some participants found them irritating while others did not find them irritating at all. Repeated phrases such as "excuse me" and naïve attention-grabbers such as "if you're not too busy" were perceived as particularly annoying.

A number of related studies have shown that a more 'humanised' output in human-computer interaction can have a positive impact on user attitudes and lead to increased rapport between the user and the computer. Peiris et al. (2000) explored different interviewing techniques in a text input/output screen-based questionnaire. Questionnaire respondents reported that they felt the 'human conversational style' which employed empathetic preambles to questions made the computer interview seem more interesting and enjoyable compared to blunt direct questioning. Furthermore, respondents answered honestly more often in the human-style condition. Research into varying system text output styles (telegraphic, fluent and anthropomorphic) revealed that the participants in the anthropomorphic condition were more than twice as likely to refer to the computer using the second pronoun 'you' and used more indirect requests and conventional politeness (Brennan and Ohaeri 1994).

The idea of treating the computer as a social entity and endowing it with emotive qualities such as politeness might be considered to be controversial given the fact that the computer does not have any real understanding about the effect its behaviour may have on its dialogue partners. Some user interface designers are opposed to the idea of anthropomorphising computers and stress that users should be discouraged from thinking that computers may have human-like abilities (Shneiderman 1989, 1993, 1998:385). This position derives from the point of view that human relationships are rarely a good model for designing effective human-computer user interfaces and that the primary goal for interface design should be predictable and controllable interaction (Shneiderman 2000)[10]. Irrespective of the stance taken for or against the notion of anthropomorphism, when the computer interface exhibits human characteristics (e.g. through animated agent gestures and natural language) the user's interpretation and perception of these need to be considered. Miller (2004) highlights this point in his article about human-computer etiquette:

> "Since a computer system will be perceived in light of the etiquette behaviours it adheres to or violates, it behoves designers to consider what etiquette our systems *should* follow or flout to elicit appropriate perceptions." (Miller 2004)

The kind of reaction that anthropomorphism elicits will primarily depend on the kinds of *expectancies* that users have or develop about the computer interface. These expectancies can be vague, caused by stereotypical classifications, hearsay or past interactions and often shape the way in which information about another person (or computer) is selected and processed (Jones 1986). Furthermore, there is also indication that expectations and social responses to computers are dependent on cultural and societal behavioural norms (Takeuchi et al. 1998; Goldstein et al. 2002); the individual's level of computer experience (Takeuchi & Katagiri 1999) and even the size of the computer device (Goldstein et al. 2002).

Social phenomena in computer interfaces may yield what is sometimes referred to as the 'black sheep' effect: perceiving imperfections in an agent with suspicion or as violating expectations (Burgoon et al. 1999; Bonito et al. 1999). Research into human-computer pragmatics has confirmed that participants evaluate the conversation differently if they are primed that one conversational partner is a computer, compared to a more naïve evaluation (Saygin & Cicekli 2002). Essentially, perceived behaviour that is at odds with prevailing expectations is likely to be amplified, positively or negatively; the black sheep effect implies that the violations are likely to be judged unfavourable or unbefitting. Subsequently, users'

[10] For a summary on positive and negative implications of anthropomorphism in computers, see for example Marakas et al. 2000.

introspective assessments of transparent social behaviour in media are likely to be more critical by nature, compared to evaluations of users' perfunctory reactions to more subtle expressions of social characteristics. It becomes crucial to achieve the appropriate level of social behaviour in the dialogue. For example, research into the impact of laughter in synthesised speech on users' perceived social bonding with the computer has shown that inappropriate type or intensity of the laugh can destroy the desired positive effect (Trouvain & Schröder 2004).

Much of the research effort into the social aspects of human-computer interaction has been focussed on the visual screen interface, which is operated by keyboard and mouse. The human-computer interaction that takes place through speech over a unimodal telephone channel is different from the visual interface, and possibly even more sensitive to linguistic and social effects (Nass & Gong 2000). The use of language in a user interface (and the use of speech in particular) is considered one of the most likely characteristics of technology that prompt a social response (Nass in Anderson 2000:95). Automated telephone services rely on speech output and the characteristics of the voice (such as the pitch, register and tone) carry sensitive information about personality and identity of the 'speaker'. For example, research comparing a number of contrasting voice personalities which ranged from 'from butler to hip youth' in a voicemail system revealed that users reacted differently to these extremes (Boyce 2000). Some participants "loved" the butler personality whereas others found "him" annoying; the voice personalities that exhibited least extreme speaker characteristics caused fewer negative reactions from users, but also resulted in fewer really strong positive reactions. Furthermore, the social interaction appears to be enforced further by the use of speech recognition technology in that it is not uncommon for users of speech-driven telephone applications to answer politely "yes please" or "no thank you" in response to system prompts. Gustafson & Bell (2000) explored how non-trained users would communicate with an animated talking agent; utterances were collected through a field experiment in which the general public interacted with the application. The utterances were analysed and almost half were categorised as examples of users socialising with the agent (e.g. "what is your name?" and "goodbye"). Nass & Gong (2000:39) state that "...social errors by the computer, regardless of the mode of output, are much more consequential for speech input compared to other forms of media."

2.5.4 Brown and Levinson's politeness theory in human-computer interaction

A handful of research papers exist which provide examples of how Brown & Levinson's taxonomy has been incorporated in the domain of human-computer dialogues. Walker et al. (1997) argued that linguistic style is a key aspect of character and proposed 'Linguistic Style Improvisation' (theory and a set of algorithms) for generating spoken utterances for artificial agents. Their work draws on Brown and Levinson's theory of linguistic social interaction in calculating the ranking of the imposition of a dialogue contribution (e.g. a request) and the face-redressive strategy employed (i.e. positive, negative, off-record). The authors propose to use the Linguistic Style Improvisation in interactive story and dialogue systems but do not report any practical application or experiments.

Similarly, André et al. (2004) propose employment of Brown and Levinson's theory for politeness behaviour in dialogue systems to endow the computer with emotional intelligence: the ability to recognise the user's emotional state and the ability to react to it appropriately. The objective of employing face threat mitigation in the dialogue is to improve the user's perception of the interaction. Their model follows Brown and Levinson's taxonomy in that politeness expressions are triggered by the relative 'seriousness' of the conversational act in *addition* to considering situational factors such as user's emotional state (in terms of valence and arousal[11]) and personality profile. This is a rather elaborate model which is both knowledge-driven (based on information about the user) and weight-driven (based on the seriousness of the face threat); the authors have yet to apply their model in an experimental dialogue setting.

Johnson et al. (2004) and Johnson & Rizzo (2004) propose the use of Brown and Levinson's politeness taxonomy for selecting different utterance types for an

[11] Examples of emotions (André et al 2004): joy (positive valence, high arousal), bliss (positive valence, low arousal), sadness (negative valence, low arousal) and anger (negative valence, high arousal).

animated tutorial agent in a virtual factory training dialogue. The motivation for adding politeness to the agent tutor is to account for the student's affective goals (motivation) and cognitive factors. They assign positive and negative politeness values to the natural language output templates. A 'direct tutor' (bald strategy) and 'polite tutor' were then compared in an experiment (Wang et al. 2005). They found that politeness can affect the student's motivational state and help them to learn difficult concepts; however, the main experiment involved only 11 participants in a between-subjects design and so the findings should be considered tentative.

McFarlane (1998) in his work on interruptions of the visual display in human-computer interaction discusses the use of Brown & Levinson's taxonomy to mitigate the negative impact of an interruption. However, his conclusion is simply that politeness is an irrelevant topic for the design of user interfaces as computers do not have 'face' and people do not have face-wants relative to their computers. MacFarlane therefore suggests that the 'bald' strategy is adequate for these purposes and should be employed.

2.6 Usability in spoken dialogue systems

The ISO 9241-11 (1998) definition of usability – originally specified for the design of visual display terminals – has broad applicability: *"The extent to which a product can be used by specific users to achieve specific goals with effectiveness, efficiency and satisfaction in a specified context of use"*. The European Telecommunications Standards Institute (ETSI) adopts the ISO standard, and enhances the definition by stating that:

> *"**Usability** is considered a pure ergonomic concept not depending on cost of providing the system. Usability together with the balance between the benefit for the user and the financial costs form the concept of **utility**. This means that an ergonomical highly usable system may have low utility for a particular user who considers the cost to be too high in relation to his or her need for using the system." (ETSI standard)*

Despite increased research into the area of SDSs and continuous deployment of new such applications into the mass-market, gaps still remain in our knowledge of usability for unimodal task-oriented systems and what it is exactly that makes users like a system (Larsen 2003; Dybkjær et al. 2004). Designing usable dialogue systems is often considered more an art than science. Indeed, several efforts have been made to define and measure 'usability' for dialogue systems, however, a unified approach has not yet been achieved. The complexity of such a process has been highlighted by the editors of the 'Journal of Natural Language Engineering', in the special issue addressing 'best practices in spoken dialogue systems engineering' (Van Kuppevelt et al. 2000):

> *"...what constitutes best practice in one design, development or evaluation situation won't be the same as what constitutes best practice in another."*

The following sections review usability guidelines and evaluation strategies which have been developed and applied in the current dialogue engineering literature to date. The literature on GUI usability is extensive, however, due to the inherent differences between graphical and telephone-based interfaces, the discussion here will focus mainly on material published on the topic of VUIs.

2.6.1 Designing for VUI usability: guidelines

Guidelines benefit from being applicable in the early stages of the design, and the degree to which the system adheres to the guidelines can be assessed in a usability inspection. However, there are a number of limitations associated with guidelines (Bevan & Macleod 1994, Grudin 1989). For example, guidelines which aim to generalise over a wide range of applications, users, tasks and environments can be difficult to interpret and their effectiveness may depend on the actual expert using them; on the other hand, more narrowly defined guidelines are likely to be applicable only for specific applications and in particular contexts. Furthermore, there is no guarantee that a specific set of guidelines will cover all relevant features of a product, nor that following the guidelines in the implementation will guarantee any particular level of usability.

A number of guidelines/principles for the design of voice and touch-tone operated telephone services have been proposed in the literature (Bernsen et al. 1997a; Bond & Camack 1999; Krahmer et al. 1997; Mohlich & Nielsen 1990; Ross et al. 2004; Marics & Engelbeck 1997; Lamel et al. 2000); all varying in terms of detail and coverage. A number of books have also been published with information useful to designing speech interfaces (Cohen et al. 2004; Kotelly 2003; McTear 2004; Weinschenk & Barker 2000). Unsurprisingly, many of the guidelines available concern the construction of system prompts. They range from broad design definitions such as "be brief" and "be consistent", to more specific definitions such as "menus should have no more than four items" and "use action-command sequence ordering in prompts" (i.e. "for the *balance* of your account, *press 1*").

The European Telecommunications Standards Institute (ETSI) has developed broad human factors guidelines for the design of information and communication technologies[12], of which a number are directly relevant to the implementation of automated telephone services (ETSI ETR 095; 096; 329; ETSI EG 201 427; 202 116; 202 076). The guidelines cover many aspects of technical development, from precise specification of hardware ergonomics such as telephone handsets and naming conventions, through to how to provide interfaces adapted for individuals with certain disabilities or special needs. Guidelines with specific reference to automated telephone services (such as prompt design, speech recognition features and menu presentations) are more general in nature.

Bernsen et al. (1996; 1997a; 1997b) and Dybkjær et al. (1996) have developed a toolset for the design of SDSs. Their work was part of the DISC project (1997-1999) which aimed to develop a detailed set of development and evaluation methods for best practices in dialogue engineering. The toolset comprises guidelines and subsumes Grice's (1975) work on the 'Cooperative Principle' (defined based on human-human interaction) which states (p45): *"Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged"*. Grice (1975) defines this further in a set of 'maxims' (while at the same time pointing out that further maxims could be added):

[12] These documents are available to download, cost-free, through the ETSI homepage: http://www.etsi.org/

QUANTITY: 1) Make your contribution as informative as is required (for the current purpose of the exchange; 2) do not make your contribution more informative than is required.

QUALITY: 1) Do not say what you believe to be false; 2) Do not say that for which you lack adequate evidence.

RELATION: Be relevant.

MANNER: 1) Avoid obscurity of expression; 2) avoid ambiguity, 3) be brief (avoid unnecessary prolixity), 4) be orderly.

The final toolset (Bernsen et al. 1997a) comprises 24 guidelines grouped under seven different aspects of dialogue (Table 1).

GENERIC GUIDELINE	SPECIFIC GUIDELINE
DIALOGUE ASPECT 1: INFORMATIVENESS	
Say enough.*	State user commitment explicitly.
Don't say too much.*	Provide immediate feedback.
DIALOGUE ASPECT 2: TRUTH AND EVIDENCE	
Don't lie.*	
Check what you will say.*	
DIALOGUE ASPECT 3: RELEVANCE	
Be relevant.*	
DIALOGUE ASPECT 4: MANNER	
Avoid obscurity.*	Ensure uniformity.
Avoid ambiguity.*	
Be brief.*	
Be orderly.*	
DIALOGUE ASPECT 5: PARTNER ASYMMETRY	
Highlight asymmetries.	State your capabilities.
	State how to interact.
DIALOGUE ASPECT 6: BACKGROUND KNOWLEDGE	
Be aware of user's background knowledge.	Be aware of user inferences.
Be aware of user expectations.	Adapt to novices and experts.
	Cover the domain.
DIALOGUE ASPECT 7: REPAIR AND CLARIFICATION	
Enable meta-communication.	Enable system repair.
	Enable inconsistency clarification.
	Enable ambiguity clarification.

Table 1. Bernsen et al. (1997a) guidelines for habitability and usability in spoken dialogues. The '*' denotes guidelines based on Grice's (1975) maxims.

Nine of the guidelines represent Grice's maxims; the remaining 15 guidelines complement these maxims by addressing idiosyncrasies which may arise when the dialogue partner is a computer, rather than a human being. In brief, the dialogue designer should ensure that the interface is uniform in manner; that 'how' to speak to the system is made transparent; that the interface provides for both novices and experts; and that problems (mainly due to issues in recogniser reliability and accuracy) can be clarified and repaired.

The usefulness of Grice's Cooperative Principle has been noted in other human-computer interaction contexts in which natural language is used to communicate with users. Some researchers have even suggested that The Cooperative Principle may be more applicable in machine-to-human dialogues than what it is in human-to-human conversation (Arons 1991).

2.6.2 Evaluating VUI usability

According to the ISO standard, usability of an interface is evaluated in terms of:

- *Effectiveness:* the extent to which the intended goals of use are achieved
- *Efficiency:* the resources (e.g. time, money or mental effort) that have to be expended to achieve intended goals
- *Satisfaction:* the extent to which the user finds the use of the product acceptable.

Evaluation can be categorised as quantitative or qualitative, subjective or objective. Quantitative evaluation is performed by quantifying some parameter and can be subjective (e.g. measuring participants' self-reported attitudes towards service usability along a scale) or it can be objective (e.g. examining system log files to establish recognition rates). Qualitative evaluation data is usually obtained through interviews with participants. The 'think-aloud' protocol is one of such qualitative analysis methods where the participant is encourage to comment about their perception of the computer interface while they use an application (mainly used in GUI evaluation). In VUIs the think-aloud protocol would interfere with the user's speaking/listening; an alternative technique to elicit how participants were reasoning during the dialogue is to carry out a 'post-verbalisation' – after the interaction has been completed (Karsenty 2001).

Studied into the relations between effectiveness, efficiency, and satisfaction for an information retrieval system and has found only weak correlations between these three aspects of usability (Frøkjær et al. 2000). The conclusion is that effectiveness, efficiency and satisfaction should be considered to represent independent aspects of usability, unless domain specific studies suggest otherwise. In the context of speech interfaces, the *effectiveness* is generally measured in terms of task success and error rates; the *efficiency* is suitably measured by examining the task completion times and the route taken to obtain the task goal in the dialogue. The *satisfaction* element is often more elusive since it usually involves a subjective measure; the core tool for obtaining information of user satisfaction is through questionnaires and interviews (Dybjkær & Bernsen 2001). The type of measure used will depend on the purpose of the evaluation and the direction of the research; in one instance it has been proposed to determine the level of usability and cost effectiveness of an automated service mainly in terms of saved agent time (Suhm & Peterson 2001).

Evaluations normally take place either in a lab or in the field and there are trade-offs with both approaches. Confounding or interfering factors can more easily be controlled for in the lab environment (e.g. for interfering noise or disruptions); on the other hand, the lab may not represent a realistic user context for the application.

Evaluations should take place throughout the design process. A number of usability inspection methods are available (Nielsen & Mack 1994). For example, cognitive walkthroughs can be performed at the early stages of a design and involve an expert, a description of the prototype of the design, a description of the task to be performed, a complete list of actions needed to complete the task and a description of the target users' knowledge and experience. In dialogue engineering, the description of the system usually takes the form of a flow-chart with details of prompts and associated processes for handling all possible user responses and all possible paths through the system, from the start of the call until the user hangs up. A cognitive walkthrough can highlight potential problem areas and where there is room for improvement.

A heuristic evaluation is generally performed by several evaluators independently to critique a system and identify potential usability problems, and usually early on in the design. For this purpose the evaluators employ a set of heuristics (guidelines or general principles) while critiquing the system. Dybjkær & Bernsen (1998) have proposed to apply the typology of cooperativity (outlined in Table 1) as a diagnostic tool for evaluating spoken dialogue systems by detecting, classifying, diagnosing and repairing user-system interaction problems. Detection of interaction problems was done by comparing expected and actual user-system exchanges in transcriptions of a train ticket booking service dialogue (Dybjkær & Bernsen 1998). They claim that the toolset has proven to work well in assisting (both trained and untrained) evaluators to identify weaknesses in dialogue systems.

Dutton et al. (1993) propose a usability research tool for assessing the attitudes towards automated telephone services of large groups of participants. The questionnaire contains a 'core' set of 20 user-perceived salient attributes which were identified in a pilot study involving observation studies, interviews with naïve users and a review of the literature (Dutton et al. 1993; Love et al. 1994). The questionnaire employs the Likert rating scale technique (Likert 1932): each attribute is represented by a short and simple statement about the service followed by set of tick-boxes along a seven-point scale, ranging form strongly agree through neutral to strongly disagree. Statements in the questionnaire are balanced, positive and negative, to counteract the response acquiescence set – the participant's natural tendency to agree during a long series of statements. Examples of two questionnaire statements are shown in Figure 2. The usability questionnaire in its entirety (as used in the experiments in the current research) is available in Appendix 1.4.

	Strongly Agree	Agree	Slightly Agree	Neither agree nor disagree	Slightly Disagree	Disagree	Strongly Disagree
Q10 I liked the voice.	□	□	□	□	□	□	□
Q11 I felt that the service was reliable.	□	□	□	□	□	□	□

Figure 2. Example of attributes for assessing users' perceived usability of automated telephone services.

Statistical evaluations of this questionnaire have proven it to be a satisfactory measurement tool of user-perceived usability in terms of reliability, validity and sensitivity (Dutton et al. 1993; Jack et al. 1993; Larsen 2003); it has been used in several experiments to assess participants' attitudes, for example, in speech vs touch tone data entry (Edwards et al. 1997), the impact of contrasting levels of speech recognition accuracy (Jack et al. 1992a, 1992b; Love et al. 1992), the use of 'beeps' to indicate when to speak in credit card number entry (Foster et al. 1992), automated banking (Larsen 1997b), and the use of metaphors for navigation in a home shopping service (Dutton et al. 1999)[13]. The questionnaire has also been subjected to factor analysis, revealing five main constructs (Love et al. 1994): quality of interface performance; cognitive effort and stress experienced by the user; user's conversational model; fluency of the experience; and transparency of the interface.

Hone & Graham (2000; 2001) provide a review of the research effort to date that has aimed at constructing a questionnaire for measuring usability. They point out that *all* techniques proposed have suffered weaknesses: 1) the content and structure is for the most part arbitrary, 2) it has not been satisfactorily validated either against subjective or objective measures, 3) there are no reports of reliability of the techniques used (test-retest and internal consistency), and 4) the data collected is used inappropriately – scores are summed or averaged without measuring that they all belong to the same construct. They propose a new usability measurement technique which they name SASSI (Subjective Assessment of Speech System Interfaces). SASSI comprises 50 Likert-style statements with 7-point scales ranging from 'strongly agree' to 'strongly disagree' (Hone & Graham 2000). It was used to evaluate eight different speech input systems yielding 226 completed questionnaires (no details on whether or not the same or different participants were used in the evaluation of each new service). Factor analysis yielded six constructs, however, only three of these achieved satisfactory levels of internal consistency measured by using Cronbach's Alpha.

[13] For further examples of how this questionnaire has been used to evaluate automated telephone interfaces see: http://spotlight.ccir.ed.ac.uk/ (last accessed 20/08/05).

Hone & Graham's critique of usability questionnaires includes that of the Dutton et al (1993) measurement tool. Some of the critique seems unfounded, such as the claimed arbitrary selection of statements, the lack of validity and inappropriate use of data. Larsen (2003) provides support for the Dutton et al. (1993) methodology after translating the questionnaire from English to Danish (cross-checked by speech and domain experts), testing it in a pilot evaluation, and then employing it in a field experiment where 310 users called a banking service. The results from that research showed that the questionnaire had satisfactory reliability (Cronbach's Alpha = 0.92) and the factor analysis corresponded well with previous research results obtained (Love et al. 1994). The SASSI questionnaire has not reported this level of internal consistency (Hone & Graham 2001). It further appears that the work on SASSI has only completed its first development iteration and then appears to have been discontinued (Larsen 2003). At this stage the SASSI questionnaire does not seem to provide any clear advantages over the Dutton et al. (1993) methodology.

The research effort into the PARADISE (PARAdigm for DIalogue System Evaluation) framework has aimed to model user satisfaction as a function of task success and dialogue cost metrics. The intention is to lead to predictive performance models for spoken language dialogue systems, allowing for direct measures of user satisfaction based on system logs and without the need for extensive experiments with participants to assess user satisfaction (Walker et al. 2000). Walker et al. (2001) have demonstrated that performance models derived via using standard metrics can account for 37% of the variance in user satisfaction (measured by a set of questionnaire statements). The objective dialogue metrics and tagging methods used in the PARADISE are reported in detail, however, the construction of the 'user satisfaction' questionnaire items is not very well documented in their reported research (Kamm et al. 1998; Walker et al. 2000); nor have they published any validation of the questionnaire measures (Larsen 2003).

Finally, recently Hartikainen et al. (2004) have proposed to use SERVQUAL – a questionnaire developed by marketing academics for the measurement of service quality – as a subjective metric for spoken dialogue system evaluation. The questionnaire (7-point scales) includes five service quality dimensions: tangibles, reliability, responsiveness, assurance and empathy, and has claimed universal applicability to any service. Hartikainen et al. (2004) employed a three-column version of the questionnaire where user attitudes are ascertained at different points: firstly, two measures of service expectations are obtained *before* the user tries the system (projected *acceptable* level and projected *desirable* level); secondly, perceived quality is measured after the user has experienced the system. These three measurements allow the system developers to assess service quality in terms of acceptance 'zones' (maximum and minimum) and compare how actual use of the service relates to these tolerance measures. Although some early findings have been reported, further development work is necessary to establish the usefulness of the SERVQUAL questionnaire for spoken dialogue systems.

The use of the SERVQUAL questionnaire for usability evaluation raises interesting points, but there appear to be some potential flaws in the assumptions they make. Considering the wording of the questionnaire items the authors propose[14], for instance, it is questionable if questionnaire items such as "service sounds like it has modern equipment" and "service gives right service at first trial" are sufficiently unambiguous and intelligible in their present form to allow them to be used as a universal and reliable tool for subjective evaluation of SDSs. Also, as pointed out by Root & Draper (1983) in their work on constructing a software evaluation tool: questionnaires are not effective for proposing features in a system which the user has no experience of using. This raises doubts whether SERVQUAL respondents can project a – hypothetical – level of "system speed" or "personal attention" before having experienced the actual application.

2.7 Summary

In the current literature, there is virtually no reference to how to design and evaluate digressive dialogues in human-computer interaction. Consequently, what is new in research presented here is the way in which it explores novel approaches to human-

[14] Extracts from the modified questionnaires used by Hartikainen et al. (2004) available via the online appendix available at: www.cs.uta.fi/hci/spi/SERVQUAL (last accessed 26/06/05)

computer spoken dialogue design by allowing the system to suspend the set turn-taking in the regular call flow and introduce new information that digresses from the immediate topic or prime goal of the call. The strategy employed in this research is multidisciplinary, bringing together research areas such as dialogue engineering, human-computer interface design, linguistic theory and usability evaluation.

Chapter 3

The challenge of an information-rich world is not only to make information available to people at any time, at any place, and in any form, but specifically to say the "right" thing at the "right" time in the "right" way.

- Professor Gerhard Fischer (2001), University of Colorado, USA -

Research methodology

3.1 Introduction

This thesis seeks to produce a body of empirical evidence detailing the impact of deploying System-Initiated Digressions (SIDs) in automated telephone service dialogues. The digressions take the form of unsolicited informational proposals aimed at notifying the caller about (new) available product or service options. The new options are not presented in the main menu of services therefore the proposal is followed by details for the caller about how to pursue the offer. The research identifies the problem space posed by the novel concept of SID behaviour and defines the tradeoffs associated with applying contrasting engineering strategies. In a series of four experiments, SIDs are implemented and evaluated by deployment into the call-flow of an already existing mass-market telephone banking service.

Central to the evaluation of digressive dialogue behaviour is feedback from the participants in the experiments – the potential users of such a system. In a real world scenario, they will come across the digressions while using the automated service to carry out their banking tasks. The aim is to create a similar realistic scenario in a controlled environment and allow participants to experience hands-on the deployment of SIDs. The research focuses on participants' attitudes (subjective data) towards the service usability and, specifically, the kind of impact that SIDs may have on their perception of the service as a whole. Objective data such as task completion and recognition accuracy are also of interest, but to a lesser degree as the purpose of the research is to design a module *within* the dialogue of an already available (and functioning) commercial application; the intention is to leave the core system flow and prompts unchanged.

3.1.1 Chapter outline

The aim of the current chapter is to provide the reader with an overview of the methodological framework and fundamental processes on which the four experiments are based. Firstly, the section for 'general objectives' provides a summary of the dialogue engineering issues and usability considerations around which the experiment studies are centred. This is followed by a detailed overview of the automated telephone banking service functionality. The flow-charts and prompts described in this section were used to implement a mirror-version of the service which was then used in the experiments. Details and some examples of the implementation procedure, the recognition engine used, development of grammars and related system settings are then presented. This is followed by a description of the design criteria – requirements capture – for the deployment of SIDs and the type of banking products that are suitably introduced through this type of dialogue behaviour.

The remaining sections describe the participant recruitment process, the experiment design and procedures. Finally, the process for collection of the data (qualitative and quantitative) and different statistical analysis methods are discussed and documented.

3.2 General objectives

The most pertinent dialogue engineering issue in the design of auditory interfaces is the careful crafting of the system output messages and instructions, which will be collectively referred to here as system prompts. The system prompts not only guide the user as to what to say and when to say it (see Section 2.3.2), but also carry information regarding the social aspects of the interaction, such as giving the system a 'personality' (see Section 2.3.1, 2.5.3). System prompts are invariably the most significant contributing factor to the user's perception of the application in auditory-only interfaces.

The SDS used in the current research is a typical examples of a mass-market automated telephone service: pre-recorded human speech is used for system output throughout the dialogue and the interaction is mainly system-directed (in the interest of robustness). System prompts in such services mainly take two forms: informational messages where the system retains the conversational turn, or prompts where the conversational turn is passed to the caller who must respond. The issue of 'how' to approach the caller, the *delivery strategy*, and whether or not to prompt the caller for a response will be addressed in the design of the system-initiated digressive dialogue.

In addition to understanding *how* to deliver system prompts to the caller, it is also necessary to decide *where* in the dialogue flow it is suitable to locate a system-initiated digression. Most system prompts in the 'core' dialogue will (as with speaker contributions in most conversations) have a well-defined point of location, such as having a greeting at the start of the interaction followed by a dialogue in which the caller identity is established and verified. In mass-market SDSs, the location of each prompt is governed by the fact that the wording and turn-taking behaviour usually remain static between each new phone call to the service. The question therefore arises of 'where' in the core dialogue it is suitable to introduce a system-initiated digression. The associated benefits and trade-offs associated with varying the *delivery location* will be addressed in the current research.

Coupled with the dialogue engineering issues 'where and how', the wording and register used in the prompts play an important role in the design of system-initiated digressions. The most relevant issues that need to be considered are: the voice (same or different voice talent to that already used in the application), prompt length, wording, register and tone. Trade-offs associated with contrasting prompt registers also need to be addressed: a system-initiated digression that integrates with the rest of the dialogue may be less intrusive, but may be less perceptible and memorable, than one that adopts a more prominent upfront interruption. Consequently, the issue of *prompt register* needs to be investigated.

These three dialogue engineering issues (the *strategy, location* and *register* of the system-initiated digression), and their impact on user attitudes towards service usability, form the core research objectives in the current research. However, these are not the only factors that have a potentially significant contribution to the user's overall perception of the service. In particular, it is expected that customer attitudes towards the system-initiated digressive proposals will vary according to the relevance of the product to the customer's specific situation. The assumption is that customers will be more favourably disposed towards the interruption caused by the digression, or more inclined to accept the product offer, if it can be deduced from their user details that they may be interested in the information. The issue of *potential need* is a further concern that will also be addressed in this research.

In order to ensure both immediate and future success of system-initiated digressive dialogue behaviour, it is vital that the callers can understand how to pursue the offer or how to obtain further information. This calls for careful consideration of the *informational content* presented in the digressive prompt along with the *cognitive ability* of the user to interpret this information and understand how to carry out the instructions. These further two cognitive issues are included in the design and evaluation of SIDs.

3.3 The core automated telephone banking service dialogue

The SLDS used in the current research was modelled on an existing real-world automated telephone banking service: PhoneBank *Express*[15]. The service provides the Bank's customers with access to personal account information (balance information or recent transactions), enables them to perform a number of banking transactions (funds transfers or ordering account statements) and amend personal details (change the Telephone Identification Number).

In order to use the service, customers must first register to obtain a personal nine-digit membership number along with a secret six-digit Telephone Identification Number (TIN) for verification. Upon registering, customers receive a membership card containing their membership number, the phone number to the automated service and a brief list of available service options along with the associated push-button for an alternative input mode. There is also a 'mini-guide' (a credit card sized user guide which, unfolded, comprises an A4 page) and a small booklet available to customers. These printed user guides contain information about system security procedures, they provide instructions on how to use the service and list available

[15] PhoneBank *Express* is a service of Lloyds TSB Bank, plc.

service options together with details of each step involved in the process of carrying out a transaction or obtaining information. The material also includes a listing of a limited number of universal commands (i.e. "cancel", "agent", "help"). It is worth pointing out that, as long as the caller has the membership number and TIN at hand, it is not necessary to have access to the membership card or mini-guide as the system prompts will guide the user through the dialogue.

Callers can choose to use either speech or push-buttons to input their responses (or employ a mixture of both). The system has some natural language capabilities in that it enables callers flexibility in how they word their requests. For example, a caller may use synonymous expressions such as "balance" and "what's in my current account", along with extraneous speech and phrases around keywords, such as in "Uhm, I'd like the..., please". The turn-taking capabilities featured in the service are limited mixed-initiative (Allen et al. 2001), which means that the system predominantly drives the conversation by the use of directed prompts, instructions and menus; however, the caller may take initiative at certain points in the dialogue and provide more information than what was requested by the system. For example, in response to the main menu "please select balance, recent transactions or another service", the caller may respond with multiple pieces of information "the balance of my current account, please", thus bypassing the 'which account' selection stage.

For the purpose of the experiments reported in this thesis, all system prompts were recorded using the same female speaker with a Southern British English accent (the same voice was also used for the recordings of the digressions). Recording sessions took place over a number of days and the speaker received voice coaching throughout in order to provide prompts congruent in pitch and speech-rate overall. Particular care was taken in order to ensure appropriate intonation of system output that requires concatenation of individual prompts, such as the generation of money amounts. In this case, several tokens of the same word were recorded with varying stable, rising and falling pitch, in order to reflect the intonation observable in phrases, such as "seventy" followed by "pounds" vs. "five".

3.3.1 Dialogue structure

The function of the dialogue manager is mainly to trigger an appropriate system prompt to be played at specific points in the call-flow; activate the speech recogniser to handle caller input; take relevant action based on the result returned by the recogniser; and perform database lookups. The current section aims to give an overview of the dialogue flow by focussing on the most prevalent components of the dialogue manager that have the biggest impact on how the caller experiences the system interface. More detailed accounts of the underlying system complexity and behaviour will only be raised where deemed necessary.

The PhoneBank *Express* dialogue consists of two main parts: a mandatory caller identification process (outlined in Figure 3) and a main menu selection stage which lists the available service options (outlined in Figure 4). Associated prompts are presented in Table 2 and Table 3 respectively. Prompts such as the SIL and REJ are generic and are played throughout the dialogue each time the speech recogniser fails to detect any speech (silence) or is unable to interpret the speech input (reject). Furthermore, a three-tiered error recovery strategy is applied whereby repeat silences or rejected inputs trigger prompts containing more detailed instructions of how to respond. Speech input is promoted in the dialogue: push-button options (DTMF) are only mentioned in the second error prompt triggered by having had two previous failed input attempts.

In the event where the caller makes three consecutive failed attempts at giving a valid input, the system enters the 'Operator Transfer' state (Figure 3 and Figure 4). The caller will first be prompted with a generic message preamble "I'm sorry, I'm having difficulty...", followed by information related to the specific dialogue stage the problem occurred in, e.g. "...with your membership number". The caller is the prompted with "It may be helpful if you are transferred to an advisor for further assistance. Would you like to be transferred?". If the caller answers "no", the system returns to the dialogue stage where the problem occurred and starts to prompt for input again. If the caller requests to be transferred the service would normally transfer the caller to a human agent. However, for the purpose of the experiment, the call transfer was replaced with the following message:

> ***"At this point you would be transferred to an agent who would help you with your enquiry. However, since this is an experiment, no agents are present. Please hang up and inform the researcher that the call has been transferred. Thank you."***

At this point it was explained to the participant that their call would, in real life, be transferred to a human agent who would be able to help them with their enquiry. The main rationale behind the decision not to transfer the caller to a human agent was to avoid introducing an additional experiment variable (i.e. noise in the research findings caused by a conversation with a party external to the application) that would be difficult to control for and that may inadvertently have an impact on user attitudes towards the automated service.

3.3.2 Identification and verification dialogue

Upon contacting the service, the caller hears an initial welcome message and must then pass through the Identification and Verification (ID&V) process in order to gain access to their account details. The system prompts the caller to give a membership number and then two digits (selected at random by the system) from the secret TIN. The caller is given a total of three attempts to give a valid membership number and matching TIN digits. A simplified flow-chart of this dialogue is presented in Figure 3 and associated prompts are listed in Table 2.

3.3.3 Main menu dialogue

Having successfully identified the caller in the ID&V process, the system proceeds with the main menu dialogue (flow-chart overview in Figure 4, prompts in Table 3). Service options are presented to callers at the main menu in the dialogue by the use of a two-tiered approach. The first half (MAIN_MENU_A) lists the most frequently requested service options; by saying "other services" the caller can access the second half of the listing (MAIN_MENU_B). All service options are active for the caller to select at either half of the menu. This means that users can volunteer input information such as "order statement" at the MAIN_MENU_A stage (before the option has been explicitly listed).

The caller requests a particular service at the main menu and, after completing the relevant sub-dialogue (e.g. a balance request or a funds transfer), is then prompted with (ANOTHER): "Would you like another service?" At this point the caller can do one of the following: respond "yes" which triggers the MAIN_MENU_A listing; immediately respond with the next service option that they require (thus bypassing the menu listings)[16]; or respond with "no" to exit the service. Before ending the phone call the system plays the message: "Thank you for calling PhoneBank *Express*. Goodbye." The caller can of course hang up at any point to end the call.

As shown in Table 3 the prompts MAIN_MENU_A and MAIN_MENU_B state that caller may use the command "help" to request more details about the available services. This option is only listed after two consecutive failures to input a valid service request and is therefore seldom, if ever[17], requested. For completeness, the system prompts triggered by the "help" command are presented in Table 3. At other stages in the dialogue, the "help" command triggers a generic form of dialogue where each help message starts with "at this point...", followed by the related third-level prompt (error=2). For example, if the caller requests "help" with giving the membership number the system output is:

> ***"At this point I need you to enter your membership number. Your membership number has nine digits. You can either say the nine digits or enter them on your telephone keypad. Please give your membership number now."***

In the experiment, it was sometimes necessary to limit caller access to certain service options. When this was the case, a SORRY message was played which included an implicit confirmation of the service option selected (e.g. "I'm sorry, the *change TIN* service is currently not available").

[16] Sometimes referred to as 'shortcut', cf. Larsen 1997a.

[17] Based on the results from participants' actual behaviour in the experiments.

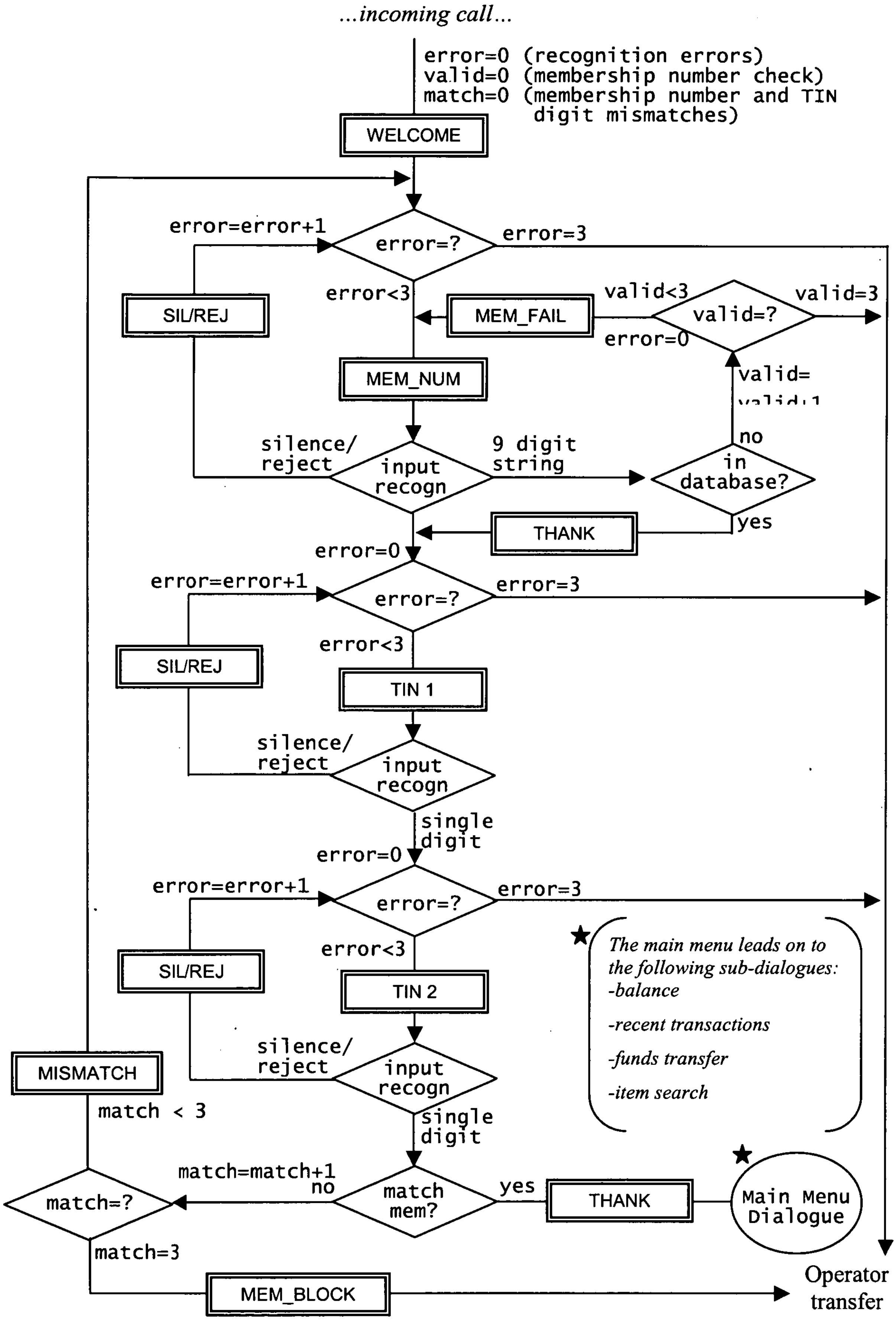


Figure 3. Outline of the identification process in the PhoneBank *Express* dialogue. Diamond shapes show system checks and recognition stages. System prompts are enclosed in rectangles.

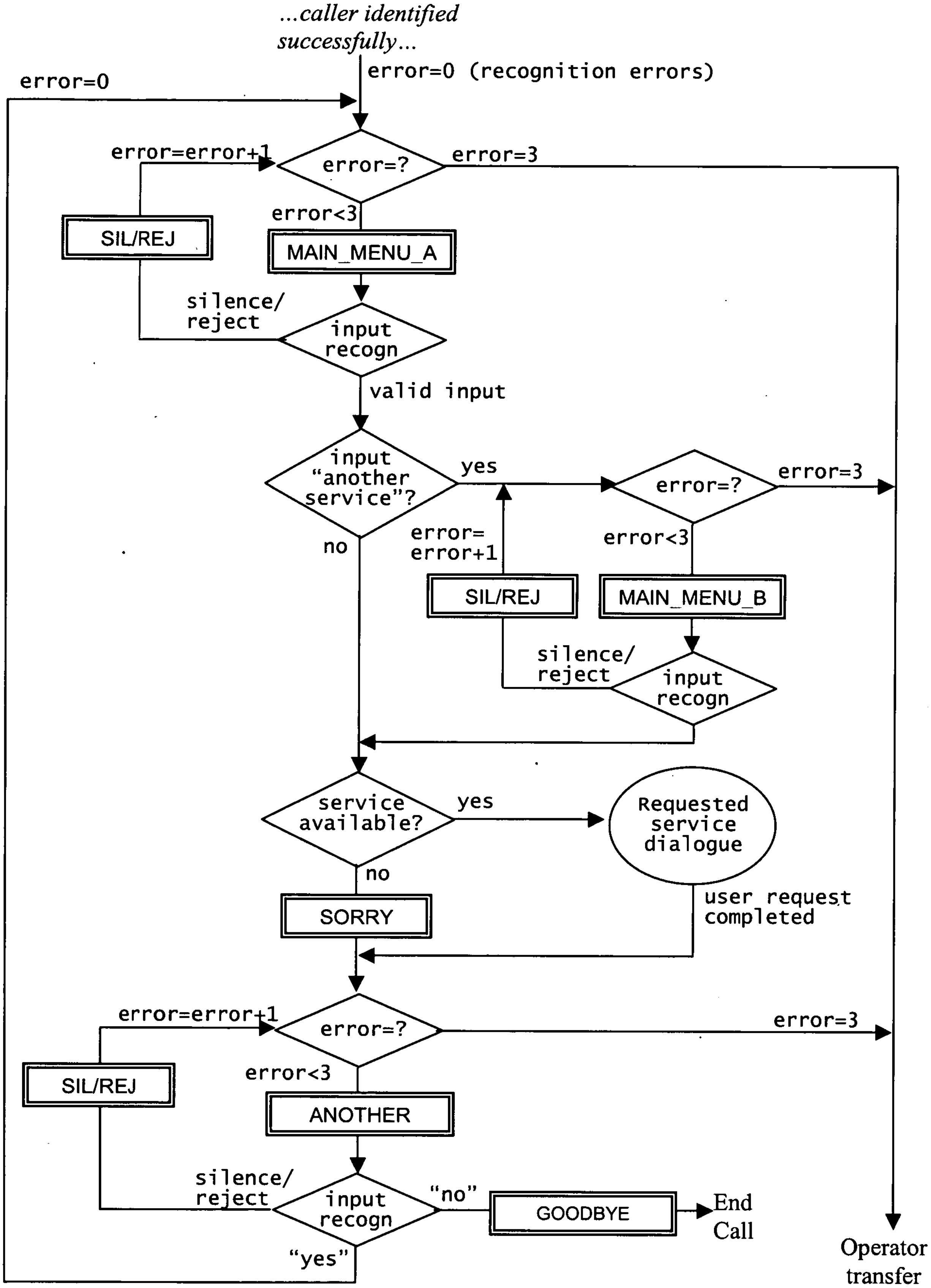


Figure 4. Outline of the main menu in the PhoneBank *Express* dialogue. Diamond shapes show system checks and recognition stages. System prompts are enclosed in rectangles.

Prompt stage	Error Level	Prompt wording
WELCOME	n/a	Welcome to PhoneBank Express.
MEM_NUM	error=0	Please give your membership number now.
	error=1	Please give your nine digit membership number now.
	error=2	Your membership number has nine digits and is printed on your membership card. You can either say the nine digits or enter them on your telephone keypad. Please give your membership number now.
MEM_FAIL	valid=1, 2	I'm sorry, that membership number doesn't match our records.
TIN 1	error=0	Please give the [X] digit of your secret TIN now.
	error=1	Please just give the [X] digit of your secret TIN now.
	error=2	You can either say the digit or enter it on your telephone keypad. Please just give the [X] digit of your secret TIN now.
TIN 2	error=0	...and the [Y] digit.
	error=1	Please just give the [Y] digit of your secret TIN now.
	error=2	You can either say the digit or enter it on your telephone keypad. Please just give the [Y] digit of your secret TIN now.
MISMATCH	match=1	I'm sorry, there seems to be a problem so I'll need to ask you for your membership number again. You can either say the digits or enter them using your telephone keypad.
	match=2	I'm sorry, this is the second time I've been unable to match your responses against our records. Please call back after checking your details.
MEM_BLOCK	match=3	I'm sorry, as this is the third time I've been unable to match your responses against our records, we're suspending your use of the service for your own security. We'll send you a new personal identification number shortly so you can call the registration line and re-register to use the service again.
SIL/REJ	silence	I'm sorry, I didn't hear anything.
	reject	I'm sorry, I didn't understand that.
THANK	n/a	Thank you.

Table 2. Prompt messages used in the identification and verification process.

Prompt stage	Error Level	Prompt wording
MAIN_MENU_A	error=0	Please select balance, recent transactions or another service.
	error=1	Please say balance, recent transactions or another service.
	error=2	You can choose from balance, recent transactions, funds transfer, item search, order statement or change TIN. Please say the name of the service you would like or say help for further details.
MAIN_MENU_B	error=0	In addition you can select funds transfer, item search, order statement, or change TIN. Which service would you like?
	error=1	Please say funds transfer, item search, order statement or change TIN.
	error=2	You can choose from balance, recent transactions, funds transfer, item search, order statement or change TIN. Please say the name of the service you would like or say help for further details.
ANOTHER	error=0	Would you like another service?
	error=1	Would you like another service?
	error=2	You can either say yes or press 1, or say no or press 9. Would you like another service?
GOODBYE		Thank you for calling PhoneBank *Express,* goodbye.
SORRY		I'm sorry the [balance; recent transactions; funds transfer; item search; order statement; change TIN] service is currently not available.
HELP	first time help requested	At this point you can get the balance on your account by saying balance, hear a list of the latest transactions by saying recent transactions, transfer money between your own accounts by saying funds transfer, search for a specific item on your account by saying item search, request an account statement through the post by saying order statement or change your secret telephone identification number by saying change TIN. Please select one of these options or say help for further details.
	second time help requested	If you would like to use your telephone keypad, for balance, press 1; for funds transfer, press 3; for order statement, press 4; for item search, press 5; for recent transactions, press 6; for change TIN, press 8. Which service would you like?

Table 3. Prompt messages used in the main menu dialogue.

3.3.4 Account selection stage

The following sections detail the service option sub-dialogues which relate to the tasks given to participants during the experiment: i.e. a balance, a search for a transaction on the account and ordering a statement. Each of these options is selected at the main menu stage. At the start of each service sub-dialogue the system first establishes the particular account the caller is interested in – i.e. the *account selection* stage. This part of the system dialogue varies according to the total number of bank accounts that the caller currently holds with the Bank. If there is only one bank account, there is no need for the system to prompt the caller to select a specific account. Also, there is no need for the system to enter the account selection stage if the caller's request at the main menu included the required account (such as in "balance of my savings account").

However, if the caller has more than one account available, and if the account has not been specified at the main menu, then the system must prompt the caller to provide the name (or number) of the required account. In the current research, all participants were provided with *two* accounts: a savings account and a current account. The account selection function simply needs to distinguish between the two accounts, and does this by asking if the request concerns the account which usually is the most frequently used, namely: "Is that for your current account?". A "yes" response sets the topical account to 'current' and a "no" response sets it to 'savings'. This is then used as input in the service sub-dialogues, described below.

3.3.5 Balance task sub-dialogue

The balance sub-dialogue with top-level prompts is presented in Figure 5. The balance sub-dialogue reads the current balance from the most recent banking business day. This can either be yesterday, as in the example presented in Figure 5, or a date and month, e.g. "at the close of business Friday the 16th of November". Any transactions on the account (e.g. cheques and debits to clear) are then taken into account and the system plays the 'projected balance' on the account. Once the balance information has been provided, the dialogue resumes at the main menu stage with "Would you like another service?".

3.3.6 Order statement task sub-dialogue

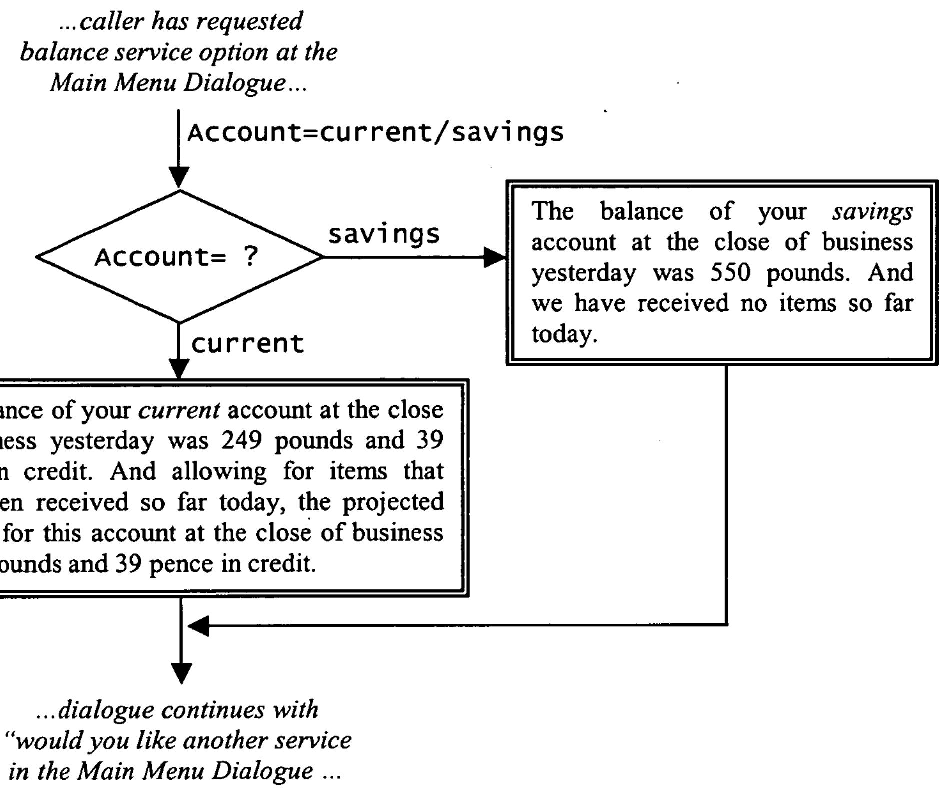
The order statement task sub-dialogue is shown in Figure 6. There is a fee charged for requesting an interim statement through the post and the system informs the caller about this and asks whether the caller wants to proceed with the statement request.

3.3.7 Transaction task sub-dialogue

One of the user tasks employed in the experiments was to find out if a cheque for £50 had been withdrawn from the current account. This task could be performed in two ways: through the 'item search' (Figure 7) service option or by requesting to hear a list of 'recent transactions' (Figure 8).

The item search option allows the caller to specify whether to search for an amount or a cheque number. Both methods are exemplified in Figure 7 which also demonstrates what happens if the search fails/succeeds or if there are more than one match for the amount specified.

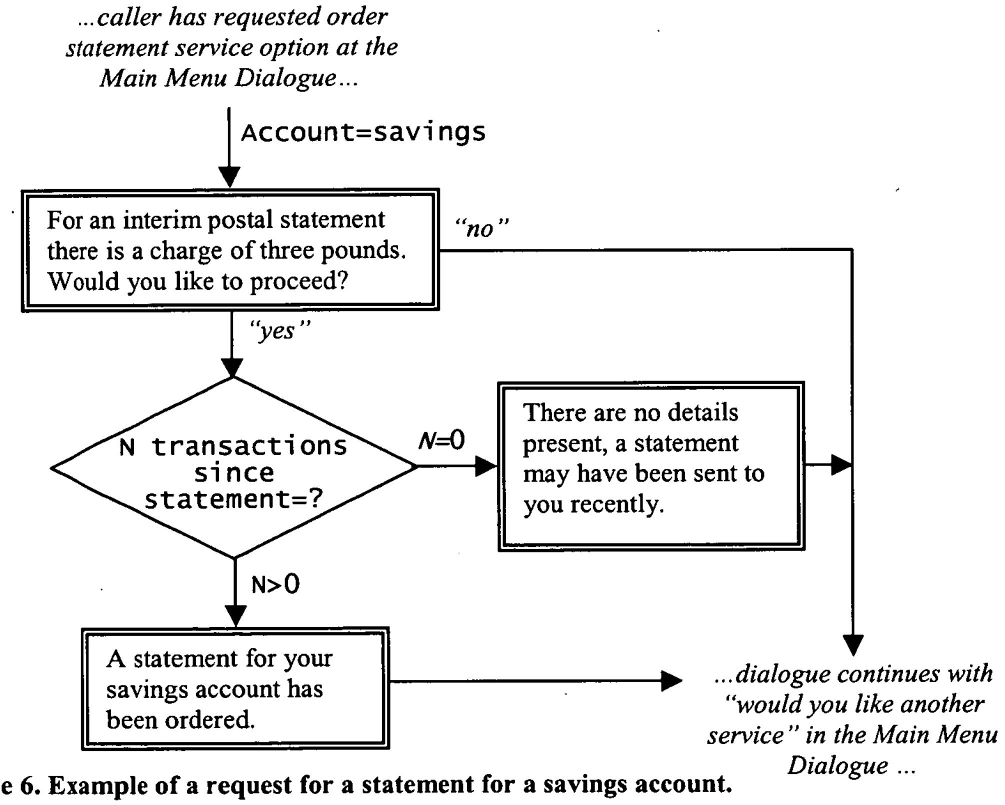
The dialogue flow for listing recent transactions is presented in Figure 8. The service lists the transactions in reversed chronological order starting with the most recent forecasted (pending) transactions and working backwards in time. Once the forecast transactions are exhausted they system plays historic items including the date that the transaction was credited/debited on the account. Transactions are read out in blocks of six and after each block the caller is asked "would you like to hear more?" until all transactions have been played (or the caller answers "no" to the questions above).


...caller has requested balance service option at the Main Menu Dialogue...
Account=current/savings
Account= ?
savings
current
...dialogue continues with "would you like another service in the Main Menu Dialogue ...

The balance of your *savings* account at the close of business yesterday was 550 pounds. And we have received no items so far today.

The balance of your *current* account at the close of business yesterday was 249 pounds and 39 pence in credit. And allowing for items that have been received so far today, the projected balance for this account at the close of business is 209 pounds and 39 pence in credit.

Figure 5. Example of a current and savings account balance request, system prompts are shown in boxes.


...caller has requested order statement service option at the Main Menu Dialogue...
Account=savings
"no"
"yes"
N transactions since statement=?
N=0
N>0
...dialogue continues with "would you like another service" in the Main Menu Dialogue ...

For an interim postal statement there is a charge of three pounds. Would you like to proceed?

There are no details present, a statement may have been sent to you recently.

A statement for your savings account has been ordered.

Figure 6. Example of a request for a statement for a savings account.

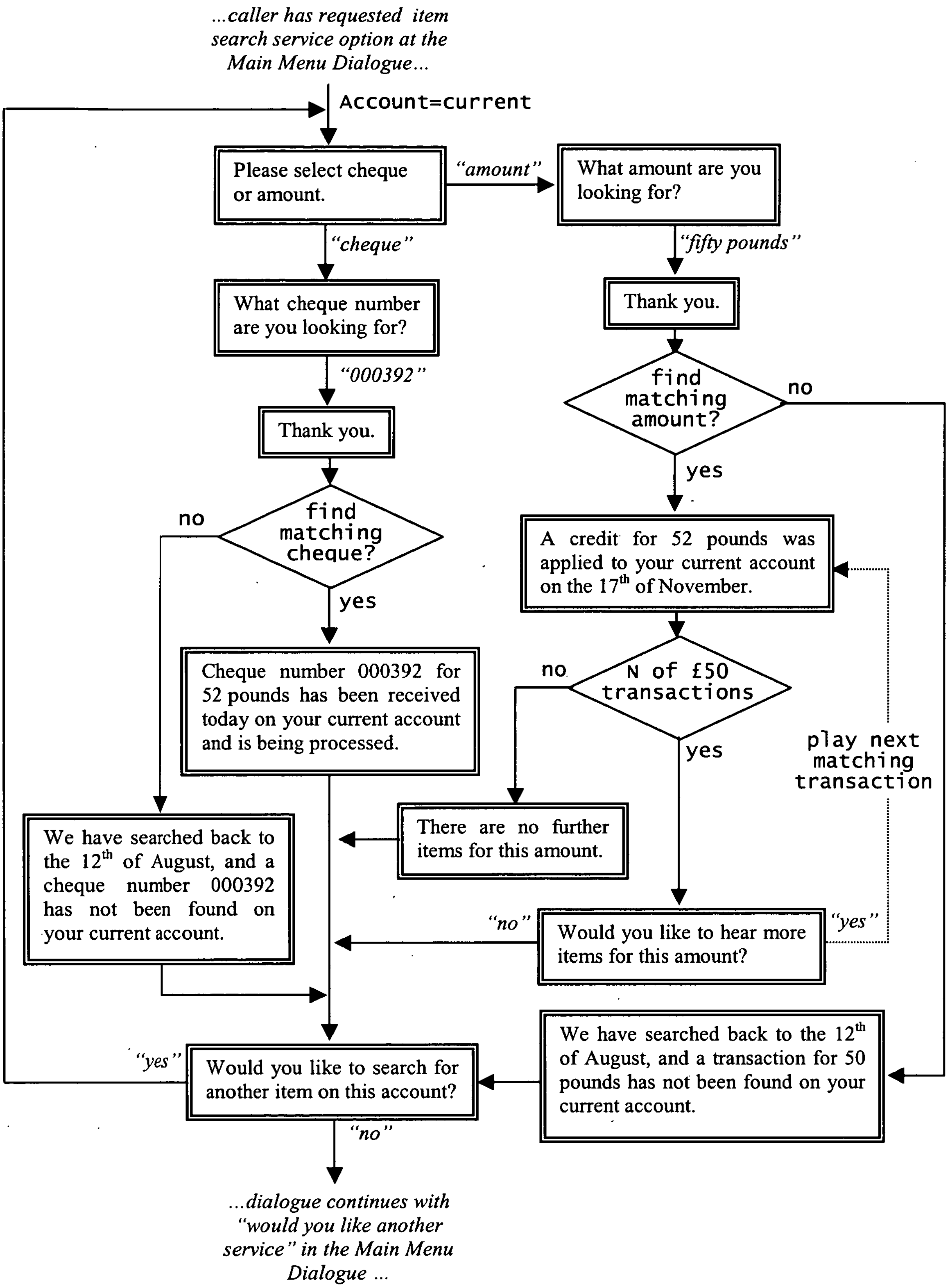

...caller has requested item search service option at the Main Menu Dialogue...
Account=current
Please select cheque or amount.
"amount"
What amount are you looking for?
"cheque"
"fifty pounds"
What cheque number are you looking for?
Thank you.
"000392"
find matching amount?
no
Thank you.
yes
no
find matching cheque?
A credit for 52 pounds was applied to your current account on the 17th of November.
yes
Cheque number 000392 for 52 pounds has been received today on your current account and is being processed.
no
N of £50 transactions
yes
play next matching transaction
There are no further items for this amount.
We have searched back to the 12th of August, and a cheque number 000392 has not been found on your current account.
"no"
Would you like to hear more items for this amount?
"yes"
We have searched back to the 12th of August, and a transaction for 50 pounds has not been found on your current account.
"yes"
Would you like to search for another item on this account?
"no"
...dialogue continues with "would you like another service" in the Main Menu Dialogue ...

Figure 7. Example of an item search request, system prompts are shown in rectangles and user responses are italicised.

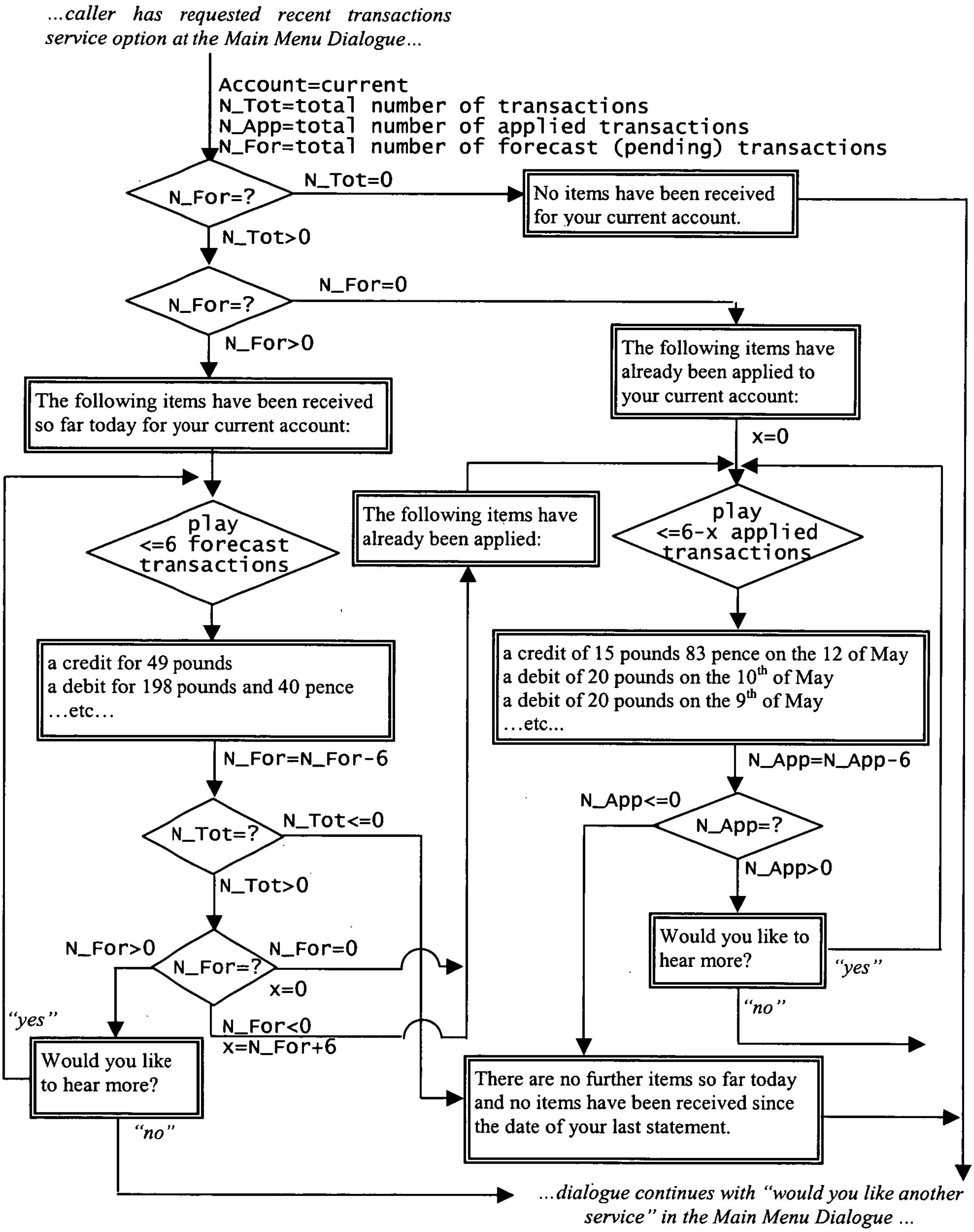

...caller has requested recent transactions service option at the Main Menu Dialogue...
Account=current
N_Tot=total number of transactions
N_App=total number of applied transactions
N_For=total number of forecast (pending) transactions
N_For=?
N_Tot=0
No items have been received for your current account.
N_Tot>0
N_For=?
N_For=0
The following items have already been applied to your current account:
N_For>0
The following items have been received so far today for your current account:
x=0
The following items have already been applied:
play <=6 forecast transactions
play <=6-x applied transactions
a credit for 49 pounds
a debit for 198 pounds and 40 pence
...etc...
a credit of 15 pounds 83 pence on the 12 of May
a debit of 20 pounds on the 10th of May
a debit of 20 pounds on the 9th of May
...etc...
N_For=N_For-6
N_App=N_App-6
N_App<=0
N_Tot=?
N_Tot<=0
N_App=?
N_App>0
N_Tot>0
Would you like to hear more?
N_For>0
N_For=?
N_For=0
x=0
"yes"
"yes"
"no"
N_For<0
x=N_For+6
Would you like to hear more?
There are no further items so far today and no items have been received since the date of your last statement.
"no"
...dialogue continues with "would you like another service" in the Main Menu Dialogue ...

Figure 8. Example of a request for recent transactions, system prompts are shown in rectangles and user responses are italicised.

3.4 Dialogue implementation

The automated banking application was implemented using the Nuance commercially available speech recognition software[18]. A general overview of the Nuance system architecture (which also applies to other similar commercial speech recognisers) is given in Section 2.2. This section gives a brief overview of the Nuance application-specific notation and components relevant to the implementation of the automated banking service in the current research.

3.4.1 Grammars

Each time a prompt is played, the speech recogniser will listen for input from the user (speech or DTMF). All allowable speech input strings (and associated DTMF options where relevant) are defined in the application's grammar file. The banking dialogue grammars were developed using Nuance's Grammar Specification Language (GSL). Each grammar in the file has a name which must include at least one uppercase character (as in the `.MainMenu` and `Balance` grammars overleaf). Top-level grammars (which are called from within the application at a specific dialogue stage) start with a full-stop (.). Sub-level grammars do not have a full-stop and are invoked only from within the grammar file itself (e.g. the grammars `Balance` and `Check` below).

The GSL notation includes brackets '( )' to denote conjunction, square brackets '[ ]' to denote disjunction and question marks '?' to denote an optional item.

In the `Balance` grammar, the input speech is interpreted by giving filler values to the slots 'command' and 'account' in `{<command balance> <account $a>}`. The '`$a`' denotes a variable which takes on the filler value given by the place holder '`Account:a`' in the grammar. If the caller specifies the account name (e.g. "The

[18] The first experiment used Nuance 6.2.1. A system upgrade (Nuance 7.0.3) was released after the completion of Experiment 1 which was then used in Experiments 2-4. New releases generally mean improvements in language models, speed and barge-in quality which in turn may have an impact on recognition accuracy. However, with regards to the grammars and vocabularies used in the dialogue implementations in the current research, the Nuance version did not significantly impact recognition performance.

balance of my current account") then this slot is filled with the value 'current'. As defined by the '?' in the grammar, specifying an account name is optional. The slot filler values are returned to the application (dialogue manager) which will in turn perform actions depending on the recognition result.

The following text is an extract from the grammar file described above:

```
.MainMenu (
      [
            Balance
            OrderStatement
            FundsTransfer
            ...
      ]
      ?please
)

Balance [
      (?Check how much ?money [(is ?there) (do i have) (have i got)]
         in my ?Account:a account)
      (?[What Check Get Want] my ?Account:a ?account
         balance)
      (?Check what in my ?Account:a account)
      (Check what my ?Account:a ?account balance is)
      (?([Want Get] a) ?(Account:a ?account) balance)
      (?([Want Get] an) account balance)
      (?([What Check Get Want] the) balance [in on for of]
         my ?Account:a account)
]     {<command balance> <account $a>}

Check [
      (?(can you) tell me)
      (WantTo know)
      (?[WantTo (can i)] [see hear check (find out)])
]
```

3.4.2 Dictionary

The Nuance software employs standard dictionary files to find pronunciations for the words specified in the grammar file (performed when the grammar file is compiled). However, some words (such as domain-specific names) are not included in the standard dictionary file and therefore need to be added into an auxiliary dictionary. The auxiliary dictionary can also include multiple entries in order to account for regional variation in pronunciation. An extract from the dictionary file includes:

```
pound p ^ u n d
pound p * U n d
overdraft o v * d r A f t
overdraft o v * r d r A f t
overdraft o v 3 d r A f t
```

The notation used in Nuance dictionaries uses the Computer Phonetic Alphabet (CPA) which provides the ability to represent the phonemes in the International Phonetic Alphabet (IPA) by using the characters on a standard computer keyboard.

3.4.3 Dialogue manager

The dialogue manager (as described in Section 2.2.3) controls the flow of the interaction: it triggers prompts to be played and performs actions on the input from the recogniser. The current dialogue manager was implemented using the Nuance Application Programming Interface (API) and the C programming language. The Nuance Dialogue Builder is a set of C functions that are used to build speech applications by creating and connecting a set of dialogue states. The Dialogue Builder provides access to all the functionality of the recognition client, such as recognition, playback for system prompts, recording and call control.

Recognition

The Dialogue Builder API functions described here gives an example of how a prompt is added to the stack of prompts to be played by the system (`AppAppendPrompt`). The relevant grammar is then set and the recognition function is called, which first plays the stacked system prompts and recognises some input.

The `AppGetRecResult` then returns a pointer to the recognition result maintained in the `App` object.

```
AppAppendPrompt(App *app, char const *prompt_name);
```

```
AppSetGrammar(App *app, const char *g);
```

```
AppRecognize(App *app);
```

```
AppGetRecResult(App *app);
```

The results from the recognition are stored in a structure and the `RecResult` functions are used to access specific results. For example, the following functions provide access to information about: the total number of answers generated by the recognition operation (used for N-best lists, such all possible matches for a membership input string and associated level of confidence returned by the recognition engine); and a transcription of the recognised utterance (string) based on the actual grammar path that was matched.

```
RecResultNumAnswers
```

```
RecResultString
```

The `RecResult` also encapsulates an `NLResult` which is the interpretation of the information bearing part of the utterance (i.e. the slots filled in the grammar). The `NLResult App` object is accessed by the following function:

```
AppGetNLResult(App *app);
```

To access an interpretation within an NL result, each slot is examined individually by using a number of functions. For example, the following functions: returns the total number of filled slots in the current interpretation; returns the name and datatype of the slot at a given index; and accesses the value of a slot for a particular datatype (in this case an integer).

```
NLGetNumberOfFilledSlots(NLResult const *nl_result, NuanceStatus int
*status);
```

```
NuanceStatus NLGetIthSlotNameAndType(NLResult *nl_result, int i,
char *buffer, int buffer_length, NLValueType *value_type);
```

```
NLGetIntSlotValue(NLResult const *nl_result, char const *slot_name,
int *int_value);
```

Parameters

The Nuance system uses a large set of parameters to control the behaviour of the various components of the recognition system and applications. The parameters can be set in multiple ways: at the initialisation of an application or process (e.g. at the command line), by using a resource file, by defining grammar contexts or at runtime through Nuance APIs dynamically changing application behaviour. The parameters associated with an application are stored in a NuanceConfig object and are identified by a module name (i.e. the Nuance software module to which the parameter applies) and a parameter name for the individual parameter used. Examples of such parameter settings are:

`ep.EndSeconds`

This parameter specifies the minimum amount of silence ('end pointing') required to indicate that the end of speech has occurred. A longer value is less likely to cut the talker off, but also makes the interaction slower.

`audio.OutputVolume`

Sets the audio output volume on a machine-independent scale of 0 to 255.

`client.RecordFilename`

Indicates the filename to use to save recognised user utterances.

`dtmf.TerminationTimeout`

Specifies the number of seconds of silence that the recognition client will allow between DTMF tones before considering the sequence complete.

3.5 The digressive dialogue – requirements capture

Section 1.3 described some motivations for introducing digressive dialogues into the dialogue of a mass-market automated telephone service. It was also established that this type of dialogue behaviour remains largely unexplored in the current literature and that it is the objective of the current research to address this issue by exploring the design and impact of digressive proposals.

In general, the task of designing new dialogues (the flow-charts and prompts) begins with a requirements capture: a process in which the purpose, criteria and functionality of the new dialogue behaviour are established. The requirements capture takes into account factors such as the client's (in this case the Bank's) needs, the characteristics of the target end user group and any technological limitations. This section provides a brief overview of the requirements capture which forms the foundation on which all digressive proposals are based. Further details about the digressive dialogues (design issues, flows and prompts) are provided in the experiment Chapters 4-7.

3.5.1 Design objectives

Ultimately, the motivational factor for introducing digressive proposals to customers in the automated service is for financial gain (albeit indirectly, and less immediately, through improved customer perception of the Bank itself). The prerequisite of this goal is, of course, that there already is an established automated banking service which customers are able – and willing – to use. At best, a digressive proposal will interest customers in obtaining a new banking product; a worst-case scenario could see customers stop using the automated service altogether after experiencing the proposal.

A key issue with respect to the design of digressive proposals is their degree of obtrusiveness: a proposal needs to be prominent enough to capture the attention of interested users but not so prominent as to impact negatively on customers' attitudes to the service (including those that are *not* interested in hearing the information). The digressive proposal should therefore aim to strike a balance between informing interested customers and avoiding intruding too heavily on the flow of the call. Secondly, the digressive dialogue should be short and concise (but informative) in order to minimise any negative impact that this may have on the caller's attitude towards the automated service.

Furthermore, the digressive proposals should only occur occasionally. This means setting a limit on how often an individual caller should be subjected to a proposal (e.g. every 90 days). Ideally, frequency of calls to the service along with information about task success rates for each caller should also be taken into account to moderate the deployment of proposals and to avoid complicating the dialogue further for callers who already may be experiencing difficulties operating the service.

From an ethical viewpoint, the proposals should employ register and contents that are not misleading and that do not attempt to coax or coerce the caller into accepting a product – especially where such acceptance entails financial commitments on behalf of the customer, such as paying interest and meeting repayments on a loan. Procedures should also be in place to make it possible for customers to opt out of hearing system-initiated proposals altogether in order to deal with callers who are strongly opposed to receiving such unsolicited information.

3.5.2 Banking products used in the proposals

Products and services particularly suited for deployment in an automated telephone service via system-initiated proposals are those which are applicable only to a subset of eligible customers, occasional promotions (new options or special offers), infrequently accessed, or pertinent to particular customers under particular circumstances (e.g. a logical link founded on some particular financial activity registered on the customer's account). These kinds of services and products would not normally be included in the design of the core automated dialogue flow, and adding these to the main menu is therefore not a viable solution. Furthermore, the fulfilment (or application) procedure of the proposed product or service should, ideally, be automated so that customer take-up can be concluded within the automated service – without having to pass the caller on to a human agent. Products with complex application procedures (e.g. mortgages) are therefore less suited for the deployment through SIDs.

Based on these criteria, two suitable banking products were selected for the experiments: an account overdraft facility and an Online Saver account. The overdraft facility may be likened with a flexible short-term loan on an eligible customer's current accounts. Each individual customer has a maximum overdraft amount allowed – a 'shadow limit' – which may or may not be disclosed by the bank at the time of handling a customer's overdraft request. The nature of the overdraft product makes it suitable for experimentation as the dialogue involved in the system-initiated proposal is straightforward and the application process can be fully automated.

The other banking product selected for this research – the Online Saver account – may be described as a newly launched product which the bank wishes to promote. The Online Saver account benefits from a preferential interest rate, but with the prerequisite that any transactions to and from the account are made through Internet or automated telephone banking only. All customers with a current account and who are also registered to use PhoneBank *Express* are eligible for opening an Online Saver. This, coupled with the fact that the process of setting up an Online Saver account may be fully automated, makes it suitable for use in experiments.

3.5.3 User characteristics

No specific target user group characteristics, which may have impact on the design of the dialogue, were identified during the requirements capture. Automated telephone banking services are generally available to members of the public and anyone eligible to open an account with the Bank is also a potential user of the automated service. In fact, most self-service automated telephone services are designed with the general public in mind and are aimed to be 'walk-up-and-use' applications that require no specialist knowledge or prior training.

3.6 Recruitment of participants

Recruitment of an appropriate participant cohort is an important part of the research and vital in terms of ensuring the reliability of the results obtained during the experiments. The primary recruitment concern in the current research is whether participants need to be banking customers and if they also need to be registered users of the automated telephone banking service (PhoneBank *Express*) in order to take part in the experiment. A possible consideration is that customer status may have a significant impact on their attitudes overall, which in turn can have an impact on their attitudes towards the automated banking service. Focussing the evaluation on the usability and functionality of the automated service, rather than customers' relationships with the Bank, should lessen this impact.

A further consideration is participants with previous experience of using the PhoneBank *Express* service (habituation effect), again which may have significant impact on their attitudes towards the service in an experimental setting. Rather than adding current personal use of the service as a variable[19] in the analysis of the data, the approach taken here was to allow each user to make a couple of 'training calls' to the service. This allows all participants to become familiar with the functionality of the service and the customer details that they are required to use during the experiment. With these facts in mind, it seems reasonable to assume that participants recruited for the experiment do not necessarily need to be Bank customers; it will be the objective of the priming material to create realistic and engaging customer scenarios for all participants.

Each of the four experiments required that a new, 'naïve' set of participants was recruited. A combination of recruitment strategies were employed: by contracting a telephone marketing company to contact members of the public; and by phoning individuals who had been sent letters by the Bank inviting them to take part in the

[19] Current use is particularly difficult to categorise. Some participants will have used PhoneBank *Express* once, some use it once a day and others use it on an irregular basis. Other participants will of course have experience from other banks' automated services which may conflict with their overall experiences and attitudes.

research. In total, 572 validated participant data sets were obtained and analysed in the experiments.

3.7 Experiment setup

The automated telephone banking application was installed on a PC with a PII 400 MHz dual processor and 256 MB of RAM, running Windows NT 4.0. A Dialogic D/300SC-E1 board (a 30-port DSP-based voice board with onboard digital E-1 ISDN telephone interface) was installed on the PC to handle the telephony connections.

Participants were seated at a desk throughout the experiment and operated the automated banking service using a standard landline telephone with push-buttons on the base of the telephone. Performance data (system log files) were collected automatically throughout the dialogue; this comprised, for each phone call, the system output (dialogue stage and error level) and the user input (speech/DTMF and recognition results). Additionally, all user utterances were stored as sound files on the computer.

3.7.1 Priming materials – persona details

Appropriate priming material is a key issue in experiment design, having significant bearing on the participant's experience of the service. In the case of PhoneBank *Express*, personal details such as membership number and a Telephone Identification Number (TIN) are required for access to the service. Each participant was given persona details to use throughout the experiment session: the name 'J. Smith', two accounts (a current account and a savings account), a membership number, a TIN and the telephone number to the automated banking service. The priming material used in the experiments is presented in Appendix 1.1.

In the interest of realism in the experiment, and in order to engage participants in the task scenarios, the account details (balance and transactional information) for the given persona were changed between phone calls to the automated service.

3.7.2 Procedure overview

In order to ensure that all participants experienced a consistent treatment overall during the experiment a carefully prepared and standardised *research script* (including a more condensed *research procedure* version) was prepared with details of what to say to the participant at each stage during the experiment. The procedure was then tested in a pilot experiment run with a handful of volunteers acting as participants. The instructions were tailored to suit each experiment; the research script and corresponding research procedure for Experiment 1 are included in Appendix 1.2 and Appendix 1.3 respectively.

On arrival, participants were greeted and then asked to take a seat at a desk containing a landline telephone (with push-buttons on the base unit), a notepad and a pen. They were then informed that they had been invited to make some phone calls to an automated telephone banking service and that they would be asked to give their opinions about the service by completing a number of questionnaires. In real life, callers would not know in advance that they were about to experience a proposal while using the service. In order to preserve this 'surprise factor' no details of the main purpose of the current research (i.e. exploring user attitudes towards the proposal) were disclosed to participants.

Details on how to use the service were kept to a minimum; before contacting the service, participants were told that they could either speak to the service or press the buttons on the telephone keypad to input their responses. Participants were also instructed that no help or assistance could be given during phone calls to the service in order to prevent such extraneous speech accidentally being picked up by the automated recogniser; further, instructions were also kept to a minimum in order to avoid inadvertently biasing participant attitudes. Each participant was then presented with a sheet of paper containing their (fictitious) persona details and the researcher went through the details together with the participant. For ethical and data protection reasons, none of the participants' personal data were used at any point in the experiment.

After receiving the appropriate priming material and instructions, all participants in the experiment made at least two practice (training) phone calls to a version of PhoneBank *Express* without SIDs. This enabled participants to become familiar with the persona details and the basic system functionality. Between each phone call to the service, participants were asked to imagine that "a few days had gone by". The account details (balance information and account transactions) were changed between phone calls in order to engage the participants in the monitoring of their accounts and to get them involved in the scenario leading up to the proposal. Participants were allowed up to three attempts to complete each phone call. A phone call was considered completed once the participant had successfully given a membership number and TIN and arrived at the menu of services in the dialogue.

Following the completion of the practice phone calls, the participants were asked to complete an attitude questionnaire to establish the reference level of the usability of the service. The procedure up to this point was comparable for all four experiments. Participants would then make additional phone calls to the PhoneBank *Express* version which featured a system-initiated digression. Following this, attitudes towards the service usability were assessed by administering further questionnaires.

More detailed accounts of the procedure employed in each individual experiment are given in Chapters 4-7. All experiments were concluded by conducting a de-briefing interview and completing a demographic questionnaire.

3.7.3 Definition and treatment of failed data sets

During the experiment session, the participant's progress was carefully monitored in order to detect any irregularities that may interfere with research findings. Problems were noted down on a specially prepared progress sheet and reasons for any hang-ups of the telephone or failed attempts at using the service were logged.

It is sometimes necessary during experiments to exclude participant data sets (questionnaires and system logs) due to certain problems or errors that arise during the session. In the current research, the reasons for excluding a participant's data set were the following:

- the participants failed – having had three attempts – to get through the compulsory identification and verification stage at the start of the call;
- the participant had not experienced a digression due to breakout before reaching this stage in the dialogue;
- the participant pursued the product/service offered in the SID immediately after having heard the offer (i.e. in the same call as the SID was deployed). In total, only a handful of the 572 participants pursued the proposal and, in the interest of data consistency, their data sets were excluded in the analysis.

3.8 Experiment design

Each of the experiment chapters in this thesis begins with the identification and documentation of the dialogue engineering objectives and issues that are explored in the experiment. Following this, a set of design criteria for the implementation and testing of contrasting SID dialogue strategies are established.

The SID represents the primary independent variable of the research, and the contrasting strategies that are tested in each experiment constitute the levels of that independent variable. For completeness and comparison, a control-group level is also included to provide an all-important baseline measurement for the absence of a SID in the dialogue (i.e. the normal version of the automated service). The other independent variables in the experiments are: gender and age group (three levels: 18-35; 36-49; and 50+). The age group distributions in the current experiments are frequently employed in social research.

The purpose of the experiment – and the approach adopted in the current research – is to manipulate the independent variables (i.e. the different levels defined) and observe the resultant effects: measured by dependent variables. In the current research, the dependent variables were responses given by participants to the individual statements in the questionnaires and answers provided in the de-briefing interview. Each experiment makes a claim (prediction or hypothesis) about the relationship between the independent and dependent variable which is tested against the null hypothesis (i.e. that there will be no differences in the dependent variable between the different levels of the independent variable).

The experiments described in the following chapters employ a combination of both between-group and within-group design methods. In the between-group design, each participant is assigned to only one experiment condition; the within-group (or repeated-measures) design lets each participant experience each experiment condition. There are trade-offs associated with each method. The within-group design is usually more economical (less participants required to obtain the same number of data entries) and generally more powerful statistically as a large part of the inter-participant variation due to individual differences is balanced out. However, the presentation order in within-group designs must be carefully balanced to compensate for possible order effects. The current research relies primarily on the between-group experiment design, as it is desirable that participants are naïve for each SID condition; once the participant has experienced a SID the purpose of the research is transparent, and the 'surprise effect' is lost.

3.9 Retrieving experiment data

Qualitative and quantitative information retrieval methods are employed in order to investigate the experiment predictions. The data are used to form conclusions relating to the deployment of SID dialogues in menu-driven automated telephone applications.

3.9.1 Quantitative data

While objective data such as recognition accuracy and navigational path can be examined by looking at system log files, subjective attitude data can only be obtained by asking participants about their opinions of the system. The primary method of quantitative information retrieval of participants' opinions of the SID involves use of attitude questionnaires. The core tool employed in the experiments for this purpose is the Likert-style usability attitude questionnaire (Dutton et al. 1993), as described in Section 2.6.2. The questionnaire comprises 20 statements and is designed to provide a subjective evaluation measure for overall service usability.

Usability questionnaires such as these are devised to provide broad measures of users' perceived service usability. Even though care and effort has been invested in devising such measurement tools, depending on the focus of the research it may be necessary to add new or more precise questionnaire items in order to obtain a more exhaustive analysis of particular interface features or modules. The approach taken in the current research is to, where necessary, complement the core questionnaire with additional items. This process requires that salient usability attitudes are identified and that care is taken when constructing the questionnaire items.

The Likert (1932; 1967) scale construction already described (section 2.6.2) is frequently used as a tool when constructing questionnaires for the measurement of attitudes. In constructing the statements for the scale Likert points out that it is important to ensure that (1932:44-46):

- all statements are expressions of *desired behaviour* and not statements of *fact* (i.e. responses should not have associated 'true or false' values);
- statements are *clear, concise* and *straight-forward* in proposition;
- stereotyping responses (i.e. the tendency to 'disagree' or 'agree' with all statements), sometimes referred to response acquiescence, should be avoided by varying the positive and negative wording of statements.

In addition, the following should be avoided in scale construction: complexity, technical terms, ambiguity, double-barrelled items, double negatives, emotive language, leading questions and invasion of privacy (Coolican 2004:177-178).

The validity of response data obtained through questionnaires is threatened by the fact that respondents may, for instance, 'gamble' with their responses (guess or be cautious about giving an honest response); have external motivation for giving a positive response (e.g. receiving payment for participating); or be careless or inattentive when filling out the questionnaire (Guilford 1967). This issue is also highlighted by Likert (1932): "The danger of not having the full cooperation of the subject cannot be overemphasized in the present promiscuous use of attitude tests". It is also important to be aware of the kinds of conclusions that may be drawn based on the data; such as, the fact that measurement of attitudes expressed by the respondent's opinions to a statement in a questionnaire does not necessarily mean a prediction of what that person would do (Thurstone 1967).

An alternative, but related, method of attitude measurement is afforded by the use of 'semantic differential scales' (Oppenheim 1992:236-241). Essentially, the scales consist of bipolar scales where extremes of adjective antonyms are used as anchors at either end (e.g. friendly/unfriendly, slow/fast). As with the Likert scale construction, the polarity should be randomised throughout the questionnaire so that there is a mixture of scales varying from negative to positive, and vice versa. Additionally, care should be taken to use terminology which is not unfamiliar to participants or difficult to interpret.

The number of response categories along the scale may vary; generally, a two-point scale will show direction of agreement, whereas a longer scale will show intensity as well as direction. Alwin (1992) investigated the level of information transmitted through scales of different lengths by comparing the number of response categories used and the resulting reliability of the attitude measurement. In sum, scale reliability was found to increase with the number of response categories. Preston & Coleman (2000) assessed the reliability, validity and discriminating power in rating scales varying widely in response categories (from 2 to 11, and a global rating score scale 1-100). They found that: (i) reliability coefficients were high for all scales and highest for scales with about 7 to 10 response categories, (ii) Cronbach's alpha coefficients were lowest for two and three-point scales and increased up to the level of 7, (iii) validity and discriminating power were statistically significant for all scales. The respondents were also asked to rate each scale on three different aspects of scale performance (from 0 to 100) for "ease of use", "quick to use" and "allowed you to express your feelings adequately"; the scale that scored best overall according to respondent preferences was the 10-point scale, closely followed by the 7-point and 9-point scales.

In sum, the 7-point scale appears to be of appropriate length for the purpose of the current research. The statements used in either Likert-scales or semantic differential scales need to be worded carefully in order to ensure that participants can respond to them confidently and accurately. The experiment scenarios need to be designed to be engaging to participants and procedures need to be in place to ascertain questionnaire respondents' full cooperation.

Objective quantifiable data are also obtained throughout the research by examining system log-files and recordings. The data collected were:

- Caller input (utterance recordings)
- Caller input mode (speech/DTMF)
- System recognition results (silence/reject/help etc.)
- Navigational path through dialogue (system prompt stage)
- Call duration
- Task completion

3.9.2 Qualitative data

Qualitative data are used throughout the research to complement findings obtained through quantitative analysis methods. The qualitative data are non-numeric and can be useful as a means for explaining relationships between findings in the numerical data and for generating new concepts. The main tool for obtaining this type of informational content in the current experiments is through a structured de-briefing interview where participants were asked a set number of questions. The wording and order of the questions in the structured interview remain the same for each respondent. The interview invites participants to speak freely about their experiences with the service and allows them to address new or different areas of interest that may have been overlooked in the design. This type of feedback from users of the automated service forms an important contribution to the overall research findings.

Care was taken in order to avoid pre-empting participant responses and questions in the interview progressed from general topics to more specific issues. For example, an initial probing question was: "Did you notice anything different with the automated service in the last phone call?" The researcher would then (for a "yes" response) ask *what* was different to elicit the respondent's awareness of the digression before moving on to more directed questions regarding this new dialogue feature in the service.

3.10 Statistical analysis – hypotheses testing

Once the questionnaire data has been collected – what useful information or conclusions can be drawn from these data? How do we determine if participants liked 'version A' or 'version B' best? How do we determine if the user rating obtained for a particular feature in 'version A' of the service is "sufficiently high"?

In the current research, the evaluation metric comprises a usability score between 1 and 7 obtained through questionnaires comprising Likert-style and semantic differential scales. Raw scores obtained through the questionnaires are polarised[20] such that a score below 4 (neutral) consistently indicates a negative user attitude with 1 as the lowest score; correspondingly, a score above 4 (with 7 as maximum) indicates a positive user attitude. Usability scores are collected from respondents who participate in the experiment (the sample population) and these data are then used to make inferences about the whole population by means of statistical modelling and testing.

The most commonly applied approach to statistical inference – and which is adopted in the current research – relies on testing the null hypothesis H_0 (i.e. the assumption that there are no differences between two versions in terms of usability). A statistical test is run on the sample data and the p value is returned which is (ranging from 0 to 1) the probability of getting the results by chance when H_0 is true. As the probability decreases, the possibility that the results are obtained by chance decreases and the

[20] The 'raw' score obtained for each questionnaire item scale runs from 1 (strongly agree) through to 7 (strongly disagree). Scales for positive statements (e.g. "The service was easy to use") are reversed such that a negative response, for example 'disagree' which has a raw score of '6', is turned into a corresponding polarised score (i.e. '2' in this particular example given).

null hypothesis can be rejected. Significance level cut-off points are adopted for the test and these in turn define the probability that the test will reject H_0 when H_0 is actually true (false positive or Type 1 error result). Conventionally, this cut-off point is set to .05 which means that there is a 1 in 20 chance that the test yields a false positive result. A *p* value below .05 is significant whereas a *p* value above .05 is non-significant (meaning H_0 cannot be rejected); *p* values below .01 are often described as *highly* significant. These two significance level cut-off points are also employed in the data analyses reported here.

The type of statistical test that is used to analyse the data set will depend on the measurement scale that has been used to collect the data. In his paper on the theory of scale measurement, Stevens (1946) defines four main scale types: nominal, ordinal, interval and ratio scales. *Nominal* data are organised into two or more discrete categories, such as 'male/female', where categories are not assumed to have any intrinsic order. Points on the *ordinal* data scale can be ordered in terms of being 'higher' or 'lower' than other points on the scale, such as 'often/seldom/never' or preference rankings. However, the actual 'size of difference' between any two points is not known. In *interval* scales the distances between points are known, such as measurement of weight and height. Finally, *ratio* scales are also defined as having an absolute zero point; thus weight fulfils the criteria of ratio scale, but not scales for measurement of temperature as the zero point in this case is arbitrary. Stevens' (1946) paper has had a strong influence on measurement theory regarding the construction of scales and the type of statistical tests that are thought to be 'permissible' for using on different scales (which have subsequently been promoted in textbooks and computer programs for statistical analyses).

Statistical analyses are categorised into parametric and non-parametric tests. Parametric tests allow for analyses of greater experimental variation (such as comparing multiple variables with multiple levels simultaneously) and have more power to detect experiment effects (Field & Hole 2003). However, parametric tests are described as having more stringent conditions attached to them compared to non-parametric tests. For parametric tests, the data collected must fulfil the property of an interval scale and, in most cases, it is assumed that the population from which the data is collected is normally distributed.

3.10.1 Non-parametric tests

Non-parametric tests do not make any assumptions about the shape of the population. In the current research, the *chi-square (Exact 2-sided)* statistical test will be applied for analyses on nominal data such as task completion rates (success/failure). This tests the null hypothesis by comparing the observed frequencies in the sample against the assumption that all the possible categories occur with equal frequency when one nominal variable is used. For *two* nominal variables (X and Y) the null hypothesis is that the relative frequencies of the values of Y are the same for all values of X. For greater accuracy, the *exact* test will be used and reported when one nominal variable is used; *Fisher's Exact Test (2-sided)* will be reported when two nominal variables are used.

For ordinal data with one categorical independent variable with two levels, the *Wilcoxon signed-rank test* is used to test within-subject differences and the *Mann-Whitney U Test* for between-subject differences. The corresponding tests used when the categorical independent variable has three or more levels are the *Friedman Test* for within-subject differences and the *Kruskal Wallis* for between-subject differences.

3.10.2 Parametric tests

The *t-test* (available for both within-subject and between-subject experiment designs) is one of the more popular tests for interval data where one categorical independent variable with two levels is used (such as when comparing usability scores for two versions of an automated service). When a variable has more than two levels, multiple *t-tests* can be performed to compare all paired combinations of the variable levels; however, this is not desirable since the risk of a false positive result is greatly increased when such multiple comparisons are performed on the same experimental data (Field 2000).

If the independent variable has more than two levels, or when it is desirable to perform simultaneous comparisons of the effects of multiple independent variables, the appropriate type of test to use is the analysis of variance (ANOVA). There may be interactions between multiple variables and variable levels that can affect the dependent variable in various ways, and the ANOVA is appropriate for testing such interactions. The null hypothesis is that the mean of the dependent variable is the same for each level of the independent variable. The ANOVA test returns the *F* value (named after R.A. Fisher who developed much of the mathematics involved) which is the ratio of two estimates of variance and is used to compute the probability values in the ANOVA. Further statistical procedures – Post Hoc tests – are required to follow up significant *F*-values for independent variables with three or more levels; these tests are used to isolate exactly where the significant difference lies.

Whether or not a significant result is obtained depends on the size of the difference between the group means; the sample size in each group (a larger sample size gives more reliable information and, if large enough, even small differences can be significant); and the variance of the dependent variable (for the same absolute difference in mean, low variance means more significant difference). Where the results of statistical tests such as *t*-tests or ANOVAs show only one or two significant differences in a set of 20 questionnaire items, these should interpreted with caution since it is statistical fact that when a number of such tests are carried out there is a high probability that at least one at the 95% level [$p = .05$] will be a false positive.

3.10.3 Notes on applying parametric tests on non-parametric data

When a rating is obtained through quantifiable means (such as the Likert or semantic differential scale), it is common practice in social research to summarise scores, to calculate the mean, examine standard deviations etc. Furthermore, parametric statistical analyses (considered more powerful and versatile than non-parametric tests) such as the *F*-test are often applied on the data in order to obtain significance values for differences in respondents' scores. However, it should be noted that such practices are fraught with controversy and have resulted in numerous debates about the type of statistical operations that can suitably be performed on Likert-style data. As the work in this thesis relies on the collection and evaluation of attitude measurements, some of these concerns will be addressed in this section, but is not intended to provide an exhaustive account of the topic.

One of the issues with Likert-style data is whether the scale, drawing on Stevens' (1946) taxonomy, should be considered interval or ordinal (Knapp 1990). The main concern is whether the distance along a conceptual scale between "agree" and "strongly agree" is perceived by respondents as being equal in size to "slightly agree" and "agree" (one of the prerequisites for an interval scale); and how to usefully interpret responses to the mid-point of the scale "neither agree nor disagree" (is it "neutral", a zero point on the scale or an indication that the respondent "doesn't know"). A further concern is that data collected through conceptual scales (especially when individual statements are analysed) are seldom normally distributed (Jamieson 2004). Such characteristics of the data would prescribe the use of nonparametric statistical tests, however, parametric tests are often favoured (and used) for such data. Some of the criticisms of non-parametric tests are presented by Yu (2002): (a) the loss of precision, (b) low power, (c) inaccuracy in multiple violations, (d) and the testing of distributions only. Taking all of the above shortcomings into account, non-parametric tests are generally not recommended.

Furthermore, some researchers have reservations about the use of Stevens' scale taxonomy. Velleman & Wilkinson (1993) question the validity of Stevens' taxonomy and state that Stevens' criteria used for selecting or recommending statistical analysis methods are inappropriate and can often be wrong: "They do not describe the attributes of real data that are essential to good statistical analysis. Nor do they provide a classification scheme appropriate for modern data analysis methods." Velleman & Wilkinson summarise some of the criticism that Stevens' work has been subjected to: [1] that restricting the choice of statistical methods to those that "exhibit the appropriate invariances for the scale type at hand" is a dangerous practice for data analysis; [2] that his taxonomy is too strict to apply to real-world data; [3] and that Stevens' proscriptions often lead to degrading data by rank ordering and unnecessarily resorting to nonparametric methods.

Labovitz (1967) provided further support of the use of parametric statistics even though certain assumptions are not met or the measurement scale is not exactly interval or ratio: [1] the insensitivity of ordinal and other nonparametric techniques (e.g. waste of information by not considering the distance between ranks); [2] the small error that results from assigning numbers to ordinal data and then treating the categories as if they conform to an interval scale; [3] test of statistical robustness which have shown that certain tests are interpretable, although selected assumptions are not met; and [4] the power-efficiency of tests. It appears that parametric tests are sufficiently robust to be applied on data that do not fully meet the criteria for the interval scale.

On the issue of non-normality of distribution and the consequences when the assumptions for the analysis of variance are not satisfied, Cochran (1947) concludes that: "the consensus ... is that no serious error is introduced by non-normality in the significance levels of the *F*-test or of the two-tailed *t*-test" but that by using the ordinary *F* and *t* tables, we tend to err in the direction of announcing too many significant results. In addition Cochran states that: "non-normality is likely to be accompanied by a loss of efficiency in the estimation of treatment effects and a corresponding loss of power in the *F*- and *t*-tests". Box (1953) further concludes: "in the commonly occurring case in which the group sizes are equal, or not very different, the analysis of variance test is affected surprisingly little by variance inequalities. Since this test is also know to be very insensitive to non-normality it would be best to accept the fact that it can be used safely under most practical conditions."

There is a controversy in the development of standardised satisfaction measurements in the form of questionnaires (and the statistical operations that can be performed on them) but that there are some advantages in using them: e.g. objectivity, replicability, quantification, economy, communication, and scientific generalisation (Lewis 2002). Even Stevens (1946) himself acknowledges that the scales most commonly and effectively used by psychologists are ordinal scales; although he states that statistics that involve means and standard deviations ought not to be used on these scales, he admits a form of 'pragmatic sanction' that "in numerous instances lead to fruitful results".

3.10.4 Notes on statistical procedures in the current research

In the current thesis, ANOVA is the primary statistical method employed using the General Linear Model (GLM) in the SPSS statistical software package (version 11.5). GLM offers four methods for computing sums of squares; Type III (the default method) is designed especially to deal with unbalanced cells in the data and will be employed in the current research. When empty cells occur in the data, the Type IV method is used[21]. The GLM computes and uses the estimated marginal (unweighted) means of the dependent variable (questionnaire scores) – not the actual observed (weighted) means. Estimated marginal means are not biased towards the cell with the largest *n*. However, actual *observed* means will be used consistently throughout this thesis when mean questionnaire scores are presented.

To follow up significant *F*-values for significant main effects for factors with more than two levels, Tukey's HSD (Honestly Significant Difference) Post Hoc test is applied throughout this thesis for between-subject factors and contrasts are used for within-subject factors. There are no Post Hoc tests for interactions between factors. ANOVAs do not provide information regarding whether specific means are significantly different from one another – simply that a significant interaction exists between the two independent factors. Where appropriate, to address whether means from interactions between factors are significant from one another, a new independent variable will be created which combines all the levels from the factors showing an interaction. For example, for an interaction between factors 'age' and 'gender' a new variable is created with six levels (males 18-35, males 36-49, males 50+, females 18-35, females 36-49, females 50+). One-way ANOVAs can then be computed to find out if the main scores across those groups (in the new variable)

[21] Information about changing the Type setting, and how to adjust this when running GLMs, is available in the SPSS software online help.

differ significantly and Post Hoc tests (Tukey HSD) can then be run to explore the significance in mean score difference between the levels.

3.10.5 Notes on the use and development of questionnaires

Throughout the experiments, the usability questionnaire described in Section 2.6.2 (and detailed in Appendix 1.4) is used as the core tool for evaluating participants' attitudes towards the service and for establishing potential changes in attitude triggered by the deployment of system-initiated digressions. The usability questionnaire has been tested for reliability and validity, and has been employed as a tool to measure participant attitudes towards automated telephone services in several experiments.

Where deemed appropriate, or necessary, additional questionnaire items will be constructed to complement the questionnaire. These additional sets of questionnaire items will focus on issues specifically relating to the wording and dialogue strategies employed in the system-initiated digressions. In creating new questionnaires there are two characteristics that need to be considered: reliability and validity (Pallant 2001).

Frequently used indicators of scale reliability are test-retest reliability and internal consistency. Test-retest reliability involves issuing the questionnaire to the same people at two different occasions and calculating the correlation between the scores (high correlation indicates a more reliable scale). The nature of the construct being measured in the current research (reactions to a SID) makes this test unsuitable as participant attitudes can usefully only be measured once. The second aspect of reliability assesses the questionnaire's internal consistency – the degree to which the items all measure the same underlying attribute. The most commonly used statistic for this is Cronbach's coefficient alpha which provides a value (ranging from 0 to 1) of the average correlation among all items on the scale. It is suitable for questionnaires with more than 10 items and a value of .75 and above is (roughly) seen as an acceptable value (Coolican 2004). Cronbach's alpha is reported in the current thesis where new questionnaires are introduced.

The three main types of validity (the degree to which a questionnaire measures what it is intended to measure) are: content validity, construct validity, and criterion validity. Content validity concerns expert evaluation of the contents of a test; in the current research this was achieved by involving five experienced dialogue engineers to critique and give feedback on proposed questionnaire items. Criterion validity concerns the relationship between scale scores and some other specified measurable criterion. The proposed questionnaire may be compared to a currently existing measure of the construct or by using the questionnaire to *predict* a relationship between events. No previous or similar measures for attitudes towards SIDs exist and therefore information about criterion validity cannot be obtained. Finally, construct validity involves the extent to which operational variables (scales) match the intended theoretical construct. Factor analysis is an example of a commonly used statistical procedure to identify underlying correlated structures in the development of scales and measures; factor analysis supports theoretical speculation, it does not 'prove' that a real psychological entity exists (Coolican 2004). In order to perform factor analysis, a large number of questionnaire data is required (Pallant (2001) refers to a suggestion of using 300+ respondent data). The usefulness of performing factor analysis has been questioned (Guttman 1977) and this statistical procedure is therefore judged to be outside the scope of the current thesis.

3.11 Summary

This chapter identified three pertinent dialogue engineering issues for the delivery of system-initiated digressions in dialogues: strategy, location and register. These three constructs form the core topics for the four experiments that are described in the following chapters. The automated telephone banking service which will be used to deploy these digressions has been described together with examples of how such voice-operated applications are implemented. The experiment method, data retrieval and statistical analysis presented in this chapter are employed throughout this thesis in order to evaluate the impact of system-initiated digressions on user attitudes.

Chapter 4

The newest computer can merely compound, at speed, the oldest problem in the relations between human beings, and in the end the communicator will be confronted with the old problem, of what to say and how to say it.

- Edward R. Murrow (1908-1965), Broadcast Journalist, USA -

Experiment 1 – Strategy for delivery (pilot study)

4.1 Introduction

No previous related studies of system-initiated digressions (SIDs) currently exist in the domain of automated telephone services; therefore, the evaluation reported in this chapter comprises a tentative pilot study to explore users' reactions towards such novel dialogue behaviour. Following the discussion in the methodology overview (Section 3.2), dialogue engineering issues pertaining to the introduction of system-initiated digressions primarily concern *where* in the call-flow to introduce the SID and the turn-taking strategies for *how* this new dialogue behaviour should be realised. This initial experiment aimed to explore in more detail the 'how' of system-initiated digressions by investigating the use of two competing strategies for delivery of overdraft information.

The other important dialogue engineering issue is, of course, to define the prompt register to be employed in the digression. The current experiment focussed on delivery strategy and, in order to avoid potentially introducing noise in the findings, the prompts in the digressive dialogue were designed with the aim of matching the tone and style of the existing PhoneBank *Express* prompt wordings and register as much as possible.

The third research objective regards an issue extraneous to the immediate dialogue engineering design concerns described above, but which forms an important factor in the overall acceptability and future success of system-initiated digressions: the caller's perceived need for the information. This in particular is likely to hold true for digressions that may be considered to have 'sales' characteristics with potential financial gain for the seller, such as the offering of an overdraft facility. Determining exactly what is or is not relevant to an individual caller at a particular moment is a complex matter involving modelling of the caller's intentions, wants, needs, motivation and goals – most of which are out of reach of the application. Potential 'logical links' can to a degree be inferred from details in the caller's bank account information but, even when SIDs are based on a complex user model, there is no guarantee that – at the time of the SID delivery – the caller actually wants or needs this additional information. Complex user modelling is outside the scope of the current research; however, taking into consideration the importance of a 'logical link' on how SID information might be perceived by the caller, an approximation of potential need will be established in the experiment by varying the participants' account balance information.

4.2 Design objectives

Three design objectives have been introduced and will be explored in the current experiment: delivery strategy, prompt register and perceived need. The rest of this section is concerned with a more detailed description of the design consideration and the implementation particulars involved.

4.2.1 Dialogue engineering objective 1: Strategy for delivery

System-initiated proposals of the type explored in the current experiment can take one of two forms, referred to here as 'Signpost' and 'Follow-on'. The Signpost strategy consists of a short informational message embedded at a specific point within the normal service dialogue to notify customers about the availability of an overdraft facility and the location of this new service option within the automated dialogue (in this case, by requesting "overdraft" at the menu of services). After the Signpost proposal information has been delivered, the system resumes the dialogue as normal. The intention behind the Signpost strategy is to interest and inform the caller without intruding too heavily on the call flow. It is then at the caller's discretion to locate and select the product option. A simplified dialogue flow-chart is shown in Figure 9 below where the overdraft proposal is deployed after the caller has obtained the balance of an account.

Potentially more intrusive, the Follow-on Strategy involves informing the caller about the availability of an overdraft facility and then prompting the caller to make a decision (and respond with "yes" or "no") to either accept or reject the overdraft offer before the dialogue can continue. If the caller responds "yes", the system launches the overdraft application sub-dialogue; a "no" response triggers a short Signpost-style message letting the customer know how to apply for the overdraft, for future reference. The system then resumes the dialogue as normal. A simplified dialogue flow-chart of this strategy is shown in Figure 10.

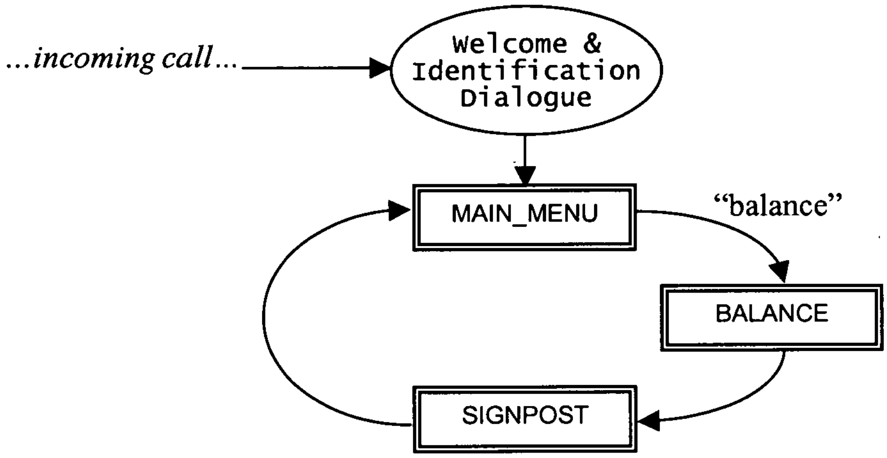


Figure 9. Overdraft proposal dialogue using the Signpost delivery strategy.

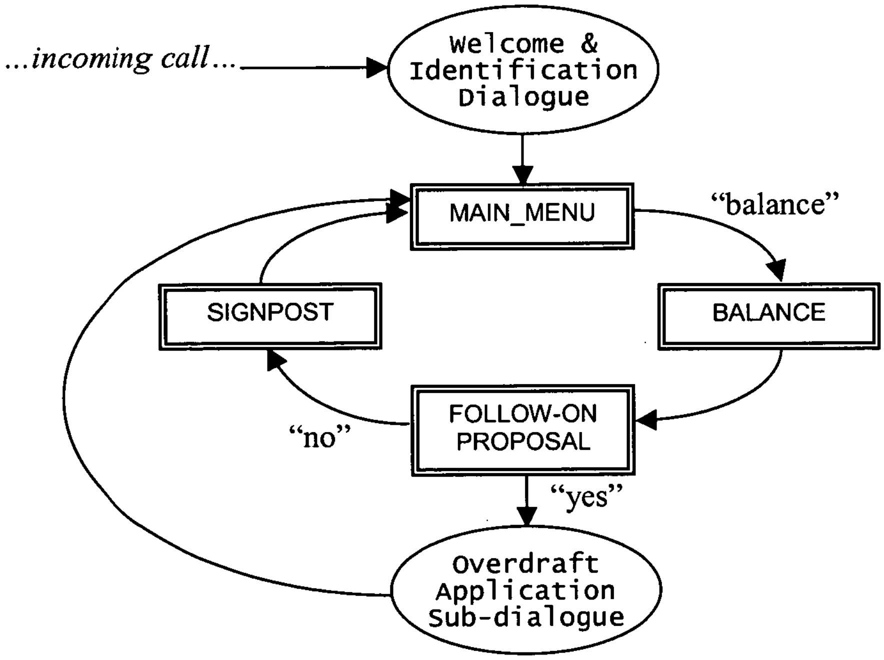


Figure 10. Overdraft proposal dialogue using the Follow-on delivery strategy.

More detailed flow-charts for the system-initiated digressive dialogue are included in Appendix 2.1.

4.2.2 Dialogue engineering objective 2: Prompt register

The wording of the Signpost prompt message was as follows:

> ***"You might like to know that you can have an overdraft of 400 pounds on this account. If you are interested, please say overdraft at the main menu."***

The wording of the Follow-on prompt message (initial prompt level) was as follows:

> ***"You might like to know that you can have an overdraft of 400 pounds on this account. Would you like this overdraft now?"***

If customer responds "no" to the Follow-on prompt...

> ***"If you would like to apply for an overdraft in the future, just say overdraft at the main menu."***

If customer responds "yes" to the Follow-on prompt...

> ***"To confirm, you would like an overdraft of 400 pounds, is that correct?"***
>
> If caller responds "yes" at this point...
>
> ***"Thank you. Your current account now has an overdraft of 400 pounds. You'll receive written information within the next few days."***
>
> Alternatively, if the caller responds "no"...
>
> ***"Your overdraft request has been cancelled."***

All system prompt scripts used in the digressions (including all error-level re-prompts) and the dialogue module for handling overdraft requests (when selected from the main menu) are detailed in Appendix 2.1. For simplicity, the overdraft application process in this initial pilot study was set to a fixed amount of 400 pounds; callers could either accept or reject this amount, there was no opportunity for negotiating a lower or higher amount.

4.2.3 Contributing factors: Explore impact of perceived need

The scenario 'potential need' vs. 'no potential need' for an overdraft was replicated in the current experiment by varying the balance of the current account. Two conditions defining the need for an overdraft were established during the experiment: one condition ensured that customers had a high account balance (of which they were aware) prior to the overdraft offer; a second condition ensured that users were aware that they had a very low account balance prior to the offer. It was thought that participants with a low balance might be more inclined to accept the overdraft or possibly find the overdraft information more helpful than those with a high balance.

The approximation of potential need by the use of simple scenarios (low/high end-balance) was selected in favour of a more complex scenario-building approach for establishing potential need among participants. There are a number of reasons for this. Firstly, participants are (as part of their task) already involved in monitoring their changing account balance and may see the logical link between a low balance and the overdraft proposal. Secondly, real-world behaviour is difficult (if at all possible) to recreate in a laboratory setting; scenarios in the form of 'persona stories' with details of purchasing behaviour, personal interests, banking activities, motivations, and so on, can be perceived as artificial by participants and can be difficult for them to interpret or to identify with. Thirdly, a persona story can only define a limited subset of a real customer's activities and, even if every participant were given a unique persona story, it would not be sufficient to account for the full range of factors that may influence real-world behaviour. Finally, participants may not be able to dispense with their own personal experiences and beliefs, making it difficult to judge whether a positive/negative response to the SID was due to the actual experiment scenario or participants' own motivational factors.

4.3 Experiment predictions

The primary aim of the experiment was to assess the relative effectiveness of the two offer strategies described above. A no-offer control group was included in the experiment design. In this way, the effect of balance level (in the absence of an overdraft proposal) could be measured as a control condition, as well as exploring the effect of presence/absence of an overdraft proposal. The main experiment predictions were as follows:

1. It was predicted that the delivery of a system-initiated digression (overdraft offer) would have a negative impact on participants' attitudes towards the automated service.
2. Being longer and requiring the user to respond, it was predicted that the Follow-on delivery strategy would have a more negative impact on participants' attitudes towards the service than would the Signpost strategy.
3. 'Potential need' for an overdraft is likely to have an impact on user attitudes towards the overdraft offer. It was predicted that the 'potential need' scenario realised through a low current account balance message prior to the overdraft offer would result in a more positive attitude towards the automated service, than would a high balance scenario.

4.4 Method

Chapter 3 provided an overview of the experiment method adopted in the current research. This section provides further details that are relevant and specific to SID Experiment 1.

4.4.1 Design

As indicated above, two different SID strategies were explored. Experiment analyses rely mainly on a between-group design; repeated-measures, within-group comparisons were indirectly achieved by running between-group analyses on the change in attitude – the differential scores – computed by subtracting questionnaire scores *before* the fourth call to the service from questionnaire scores *after* the fourth call. Grouping (between-group) variables included age, gender, offer type (No-offer, Signpost and Follow-on) and final account balance (high or low).

4.4.2 Participants

A total of 168 participants (73 males and 95 females) contributed to the evaluation in Experiment 1. Participants were recruited from the general public and only two of them had no previous experience of using PhoneBank *Express* (a further 64 participants had experience of other automated telephone banking services). One advantage with using naïve users was that factors such as prior habituation effects and experience with the PhoneBank *Express* service could be monitored and controlled more easily during the experiment.

4.4.3 Materials

Participants were given the following personal banking details (described further in section 3.7.2): a membership number, a TIN, details of two accounts (one savings account and one current account). In addition, participants were given a copy of the PhoneBank *Express* 'mini-guide'. In the third error-level prompt for capturing caller membership number the system made reference to "the membership number as printed on the membership card" (3.3.2) and participants were also provided with a membership card.

4.4.4 Procedure

Participants were assigned to one of the six experiment conditions at random: No-proposal control group high end-balance; No-proposal control group low end-balance; Signpost proposal group high end-balance; Signpost proposal group low end-balance; Follow-on proposal group high end-balance; or Follow-on proposal group low end-balance. Upon arrival, the participant was greeted and asked to take a seat by the telephone. A questionnaire was completed to obtain demographic and technographic information about the participant (Appendix 2.3).

The participant received instructions about the experiment and was then given a sheet containing the fictitious persona details. The participants' tasks were to make telephone calls (four in total) to the automated banking service to find out the balance of 'their' current account and then to find out if a cheque for £50 had been withdrawn from 'their' current account. Participants were asked to take a note of the balance and the result of the cheque search (a copy of the task sheet can be found in Appendix 2.2). These two tasks were then repeated through each of the four phone calls. On the fourth call participants (except those in the control group) were exposed to an overdraft proposal; the control group participants used the standard, proposal-free, version of the service in their fourth call.

The balance and account transaction details changed between phone calls in order to engage the participants in the monitoring of the accounts and to involve them in the scenario leading up to the proposal. For half of the participant group the balance would decrease between calls to the service nearing zero in the fourth (final) call reflecting the 'potential need' scenario (Table 4). For the rest of participants the balance fluctuated but remained at a similar level.

	Phone call 1	Phone call 2	Phone call 3	Phone call 4
'Potential need' Low end balance	£386.50	£280.07	£158.90	£4.65
'No potential need' High end balance	£486.50	£380.07	£320.07	£375.07

Table 4. Balance information. Participants make four phone calls in total (progressing from phone call 1 to 4). SIDs were deployed in the final phone call 4.

The balance information is available by selecting "balance" at the main menu of services; the search for the cheque transaction can either be completed by asking to hear "recent transactions" or by requesting the "item search" option at the menu. The dialogue functionality for a balance and item search task is explained in further detail in sections 3.3.5 and 3.3.7.

The experiment session proceeded in a number of clearly defined stages which are outlined in Table 5 below. All participants in the experiment made their first three phone calls to the same, core, version of the service (PhoneBank *Express* without proposals). This enabled participants to become familiar with their persona details and the service functionality. Following the completion of the third call, participants were then asked to complete an attitude questionnaire (Appendix 1.4) to establish the reference level of the usability of the service; this questionnaire will be referred to here as 'UQ0'.

Experiment stage	Experiment condition	Materials used
Welcome, introduction, priming	Same for all participants	Demographic questionnaire Persona details Membership card Service 'mini-guide'
Three phone calls to core service	Same for all participants	Task sheets (in each call obtain balance; find £50 cheque)
Usability assessment	Same for all participants	Usability questionnaire (UQ0)
One phone call to service with proposal 6 versions implemented	1: No-proposal, High balance	Task sheet (obtain balance; find £50 cheque)
	2: No-proposal, Low balance	
	3: Signpost proposal, High balance	
	4: Signpost proposal, Low balance	
	5: Follow-on proposal, High balance	
	6: Follow-on proposal, Low balance	
Usability assessment	Same for all participants	Usability questionnaire (UQ1)
De-briefing interview	Same for all participants	De-briefing interview

Table 5. Overview of Experiment 1 procedure.

Title	Experiment 1: strategies for delivery	
Design		One independent sample, between-subjects design adopted
Predictions	E1.1	The system-initiated digression would have a negative impact on participant attitudes to service usability.
	E1.2	The follow-on delivery strategy would be rated more negatively than would the signpost strategy in terms of service usability.
	E1.3	The digression would be more positively received by participants with a low current account ('potential need') balance than by participants with a high balance.
Independent variables	1	Application: service version (3 levels)
	2	Application: account balance (2 levels)
	3	Participant: gender (2 levels)
	4	Participant: age group (3 levels)
Dependent variables	1	Usability questionnaire, 'UQ0' and 'UQ1' (1-7 Likert scale)
Other data		De-briefing interview
Location		University Research Centre, central Edinburgh
Participant cohort		*N* = 180 (target, 30 participants in each experiment condition)
Remuneration		£10
Duration		Approximately 35 minutes

Table 6. Summary table of the SID strategy Experiment 1.

During the fourth and final call to the service (except for those in the control group), participants experienced the overdraft proposal following the current account balance enquiry. After this phone call participants completed the same attitude questionnaire but focussing on their experience of the service in last call (referred to as 'UQ1'). The session was ended with a de-briefing interview (Appendix 2.4). A summary of the experiment design is provided in Table 6 above.

4.5 Results

The results analysis presented in this section was based on data entries from participants (N = 168) who had managed to successfully complete all their four phone calls to the service. Results include: demographic/technographic details, speech recogniser performance, task completion rates, usability ratings and de-briefing interview data.

4.5.1 Demographic/technographic data

Table 7 details the participant age and gender distribution for each experiment condition. The sample was overall well balanced by gender, although some cells were slightly over represented. There was also a bias evident towards the youngest age group, consequential from the recruitment process.

		Experiment condition						
		No-proposal control group		**Signpost proposal group**		**Follow-on proposal group**		
Age group	**Gender**	**Low balance**	**High balance**	**Low balance**	**High balance**	**Low balance**	**High balance**	**Total**
18-35 years	**Male**	2	5	8	4	8	4	31
	Female	9	8	9	9	7	7	49
36-49 years	**Male**	6	4	2	3	3	5	23
	Female	2	4	4	6	4	5	25
50+ years	**Male**	3	3	3	3	4	3	19
	Female	5	3	4	3	4	2	21
Total		27	27	30	28	30	26	N = 168

Table 7. Analysis of participant cohort by age, gender and experiment condition.

About 39.3% of participants (N = 66) stated that they had used an automated telephone banking service for their personal banking needs, prior to taking part in the experiment (Figure 11). The use of automated telephone banking was particularly prevalent in the younger age group (18-35 year olds) and among females in the mid age group (36-49 year olds).

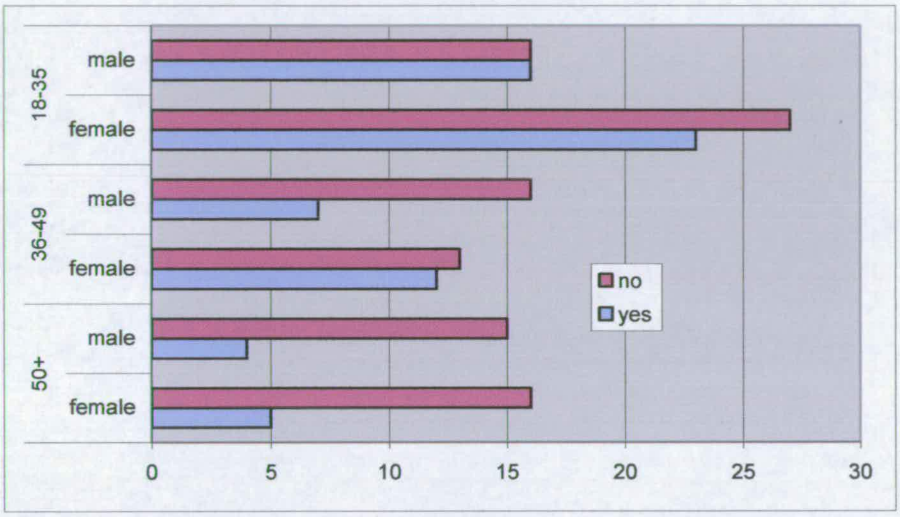


Figure 11. Use of an automated telephone service for participants' personal banking.

Participants who had used automated telephone banking were also asked which service they used, how often they used the service, if they used speech or touch-button/DTMF, and what type of banking transaction they performed using the service. Automated telephone banking services most frequently featured (with 2+ responses) amongst participants are presented in Table 8. Bank of Scotland and the Royal Bank of Scotland were the two most frequently used automated banking services in the participant sample. Only two participants had experience of using the application featured in the experimental research (PhoneBank *Express*).

Bank of Scotland	22	Lloyds TSB (PhoneBank *Express*)	2
Royal Bank of Scotland	15	Nationwide	2
Halifax	6	Barclays	2
Alliance & Leicester	3	Clydesdale Bank	2
Natwest	3	MBNA	2
HSBC	3		

Table 8. Most frequently used automated services.

When asked how they operated these automated telephone services (Table 8), the majority of participants (71.2%) stated that they gave responses by pressing buttons on the telephone keypad, while 12.1% used speech and 15.2% claimed that they used both speech and touch-buttons. One participant could not remember which input mode they had used when operating the banking service.

Frequency of use and types of banking transactions performed through the telephone banking service are presented in Table 9.

once a day	1
2+ times a week	10
once a week	7
2+ times a month	13
once a month	16
less than once a month	13
once a year/rarely	3
never	3

balance enquiry	35
funds transfers	27
pay bills	20
hear recent transactions	17

Table 9. Frequency of use of automated telephone banking and most frequently used services.

4.5.2 System recognition performance

The system log registered a total of 9,067 user inputs during the course of the experiment: 4,783 (52.8%) of these were through speech, 3,746 (41.3%) through touch-button presses and 538 (5.9%) were silences. Of the total number of touch-button inputs, 139 (3.9%) were incorrect button presses that had been rejected by the system.

The quality and performance of the recognition engine is often claimed to have significant bearing on users' attitudes towards the service. An error-prone automated service that repeatedly mis-recognises user utterances and requires users to repeat themselves is likely to be perceived to be both annoying and unreliable. To assess the performance of the speech recognition engine used in the current experiment, the 4,783 user speech inputs were first hand-transcribed and then compared[22] with the

[22] The Nuance Developer Software provides a tool for automating this process.

string representations defined by the grammars and dictionaries. Following this procedure, the transcriptions could be grouped into in-grammar (*N* = 4,323 or 90.4%) and out-of-grammar (*N* = 460 or 9.6%) user input.

A breakdown of recognition results is provided in Table 10. Of all in-grammar utterances, 94.3% had been recognised accurately by the system, 1.5% had been misrecognised (interpreted as something else) and 4.1% had been rejected. Of the out-of-grammar utterances, 22.2% were misrecognised by the system (incorrectly interpreted as valid input string) and 77.8% appropriately rejected.

	RECOGNISED	MISRECOGNISED	REJECTED	TOTAL
in-grammar	4,078 (94.3%)	67 (1.5%)	178 (4.1%)	4,323
out-of-grammar		102 (22.2%)	358 (77.8%)	460
TOTAL	4,078	169	536	4,783

Table 10. Speech recogniser performance.

The core recognition engine (grammars, dictionaries and language models), the core dialogue and system prompts change very little from one experiment to the other. Subsequently, participant input responses and the accuracy of the recogniser do not vary significantly throughout the current research. No further analyses of the recognition performance will therefore be presented in Experiments 2-4 chapters; the primary focus of the research hereon is on task completion and usability evaluation.

4.5.3 Task completion

Task completion rates were based on system log data and required that: (1) the participant had taken a (correct) note of the current account balance; (2) they had obtained information that the cheque for £50 had cleared on their current account. Overall, task completion rates for both the balance request and cheque transaction were high (>90%, Table 11). Participants were not required to apply for an overdraft in call three; two participants did, however, apply for an overdraft and their data has been excluded from all further analyses.

	Call 1	Call 2	Call 3	Call 4
Balance	164 (97.6%)	163 (97.0%)	167 (99.4%)	168 (100%)
Cheque transaction	155 (92.3%)	162 (99.4%)	158 (94.0%)	162 (96.4%)

Table 11. Task completion success rates for each of the four phone calls to the automated service.

4.5.4 Usability ratings prior to experiencing the SID (UQ0)

In the fourth phone call to the service, participants experienced a system-initiated overdraft offer following the balance request. Participants' attitudes towards the service were measured both following their third practice phone call (UQ0), immediately prior to experiencing the SID, and after completing the phone call with the SID delivery (UQ1). Responses to the usability questionnaires were analysed, both in terms of overall mean scores and according to means for individual attributes (per statement analysis).

A univariate ANOVA was run on participant responses from UQ0 with age, gender and balance (two levels: higher/lower) as between-group variables (Table 12). No significant main effects for age, gender or balance level were revealed. There were, however, significant interactions for the factors age and balance level (Figure 12) and between the factors gender and balance level (Figure 13).

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Between-participant effects					
AGE	4.518	2	2.259	2.830	.062
GENDER	.183	1	.183	.229	.633
BALANCE	.054	1	.054	.067	.795
AGE * GENDER	.532	2	.266	.333	.717
AGE * BALANCE	8.628	2	4.314	5.404	**.005**
GENDER * BALANCE	7.662	1	7.662	9.598	**.002**
AGE * GENDER * BALANCE	1.365	2	.682	.855	.427
Error	124.535	156	.798		

Table 12. ANOVA on overall usability mean scores (UQ0), all participants (*N* = 168).

There are no *Post Hoc* tests for interactions; instead, two new independent variables were created for age and balance combined (six levels); and gender and balance combined (four levels). A one-way ANOVA was computed to investigate if the means differed significantly within each of the groups. The ANOVA revealed significant overall differences for the age and balance variable [*df* = 5, *F* = 3.660, *p* = .004]; the younger age group (18-35) with a high balance was significantly more positive towards the usability of the service overall than both the mid age group (36-49) with a low balance [*p* = .006] and the older age group (50+) with a high balance [*p* = .031].

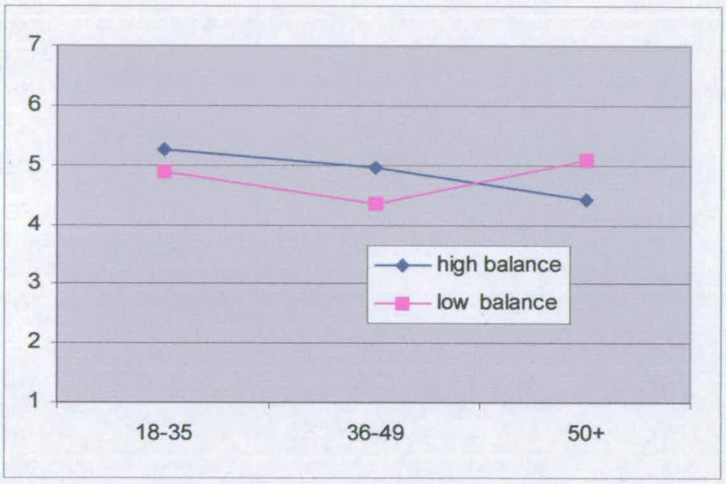


Figure 12. Overall mean usability scores (UQ0) grouped according to balance level and age, all participants (*N* = 168).

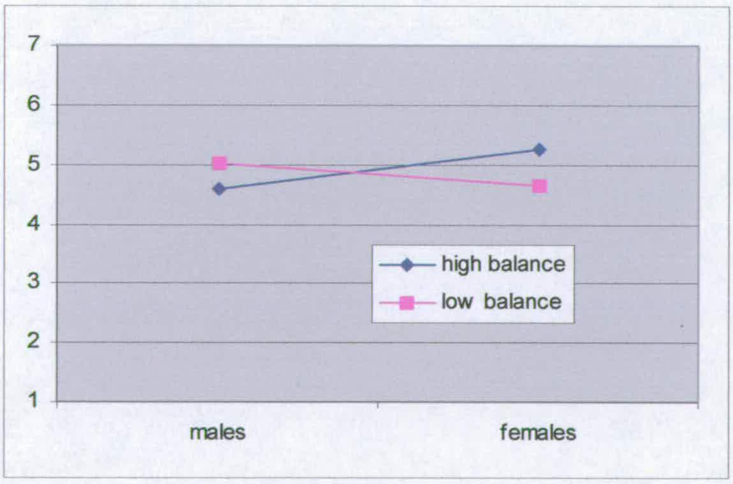


Figure 13. Overall mean usability scores (UQ0) grouped according to balance level and gender, all participants (*N* = 168).

In the gender/balance group variable, the overall difference was significant [df = 3, F = 5.050, p = .002]; female participants with a high balance were significantly more positive to the service compared to males with a high balance [p = .008] and compared to females with a low balance [p = .008].

Univariate ANOVAs (with between-group variables age, gender and balance) were also performed on each of the individual 20 UQ0 statements; score profiles for main factors are shown in Chart 1, Chart 2 and Chart 3. There were moderate significant main effects for age (Chart 1) with regards to *feeling under stress* [df = 2, F = 3.371, p = .037], perceived *politeness* [df = 2, F = 3.202, p = .043] and *level of concentration* needed when operating the service [df = 2, F = 4.144, p = .018]; however, *Post Hoc* tests revealed no further significances between the three age groups for either of these three usability attributes. There was also significant results for age on the *perception that the service was too fast* [df = 2, F = 4.311, p = .015] where *Post Hoc* tests showed that the youngest age group (18-35) were significantly more positive towards the speed of the service compared to the oldest age group (50+) [p = .015]. The *feeling of being in control* of the service also showed statistical significance [df = 2, F = 4.585, p = .012] with the younger (18-35) age group taking a significantly (Post Hoc) more positive attitude than the mid-range (36-49) age group [p = .039].

Chart 2 shows means and main effects for gender. Female participants rated the service *politeness* more positively than male participant [df = 1, F = 9.120, p = .003]. Male participants on the other hand were more positive than females with regards to the *level of concentration* needed to operate the automated service [df = 1, F = 5.355, p = .022].

Chart 3 above shows mean scores based on the balance level that participants experienced during their phone calls to the service; ANOVAs revealed no statistically significant differences between the decreasing/low balance and the static/higher balance.

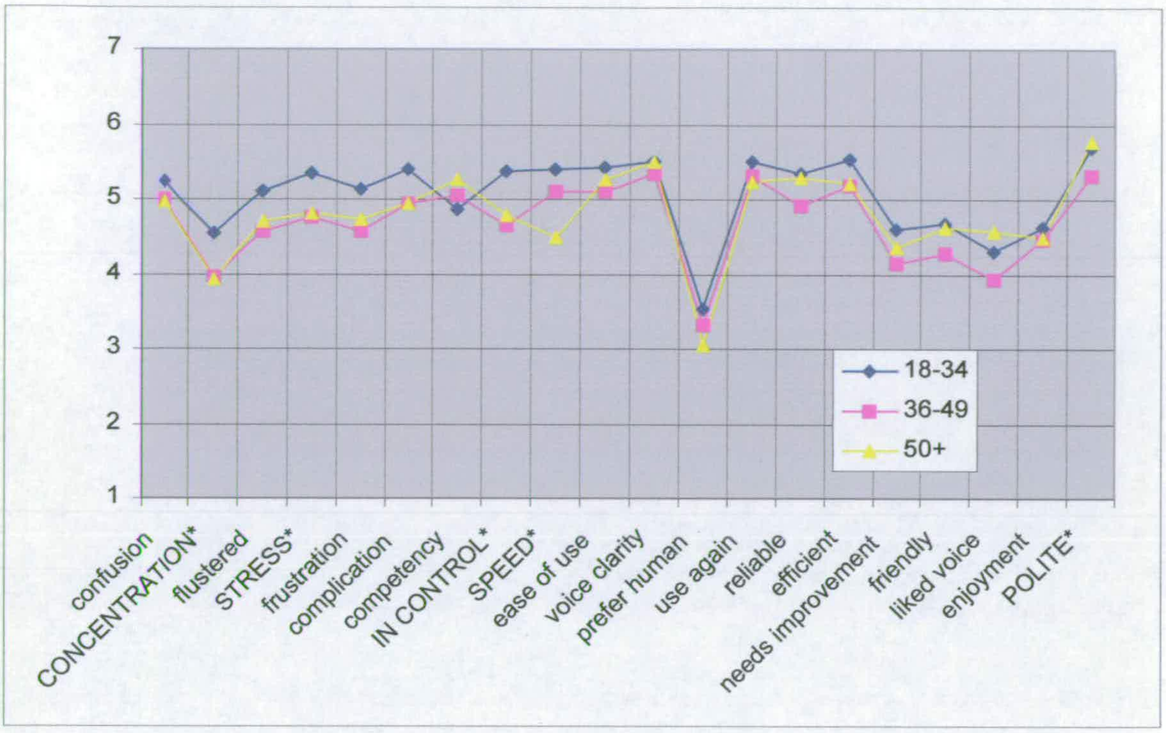


Chart 1. Main scores for UQ0 attributes, split according to age factor with three levels, all participants ($N = 168$). Statistically significant items have been capitalised and the two levels of significance are marked by using stars (*p<.05, **p<.01).

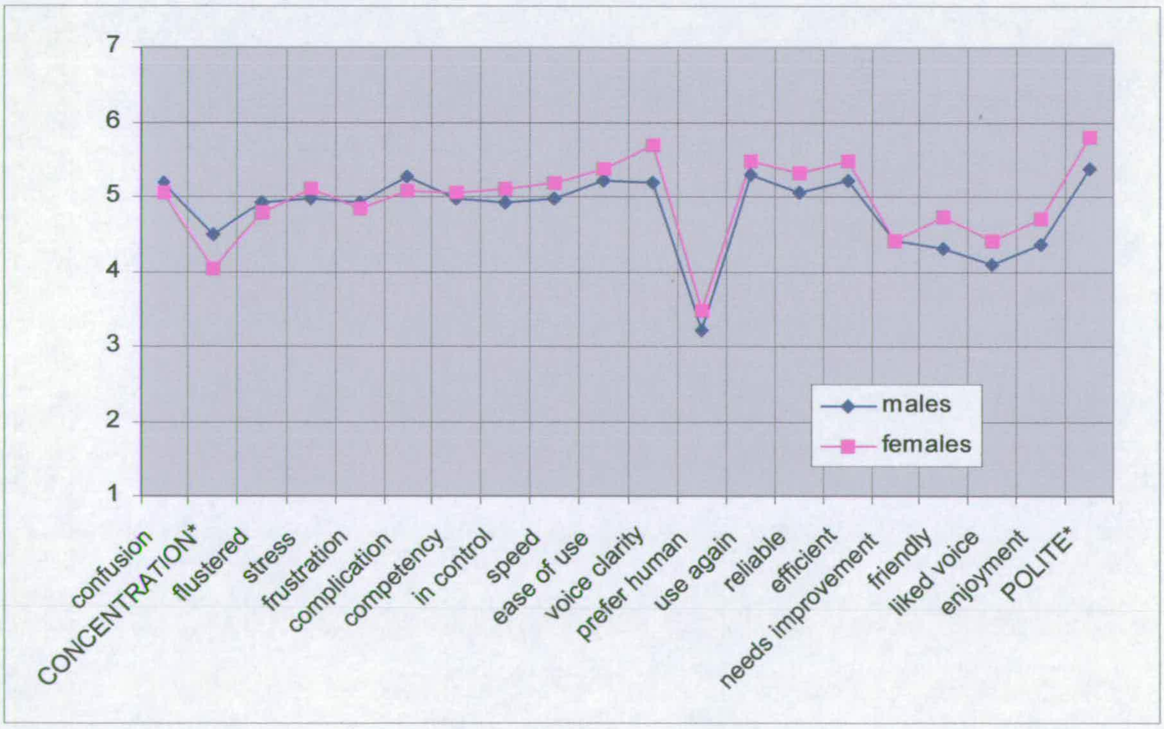


Chart 2. Main scores for UQ0 attributes, split according to gender [*p<.05; **p<.01], all participants ($N = 168$).

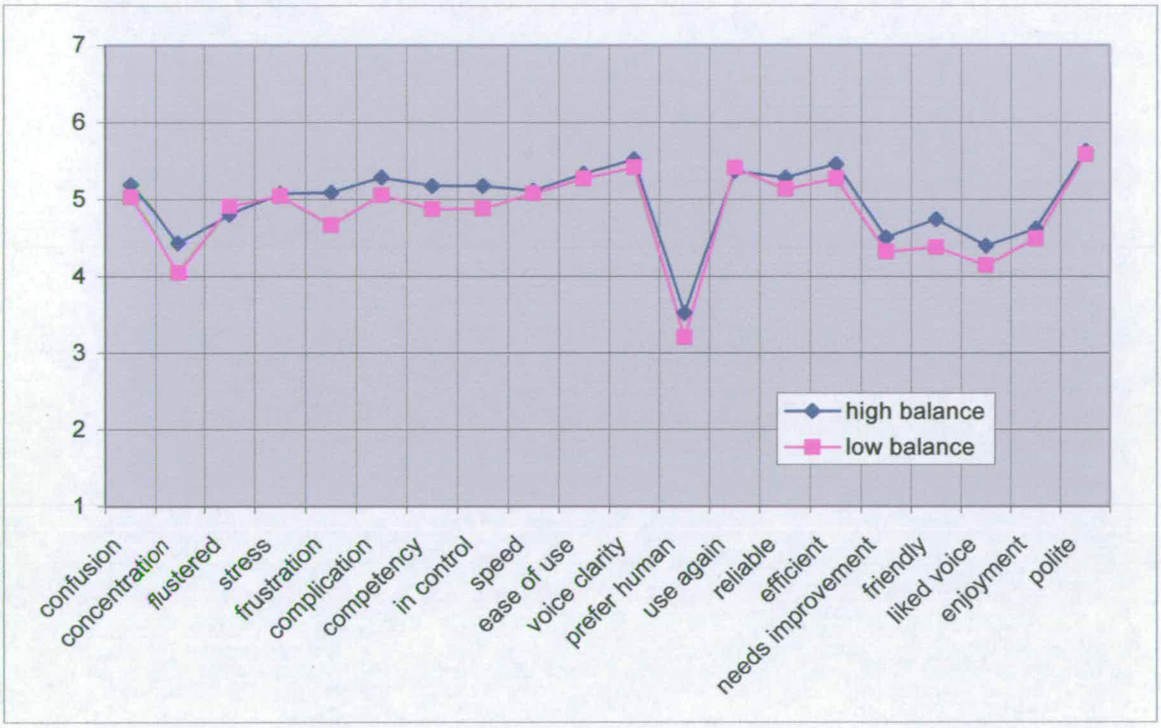


Chart 3. Main scores for UQ0 attributes, split according to balance level, all participants (*N* = 168).

Analyses of the overall questionnaire means in Figure 12 and Figure 13 above revealed significant interactions between the factors age and balance level; and between the factors gender and balance level. There was further evidence of these interactions in the analyses of individual questionnaire attributes. These findings are summarised in Table 13; the direction of the interaction follows the same patterns as in Figure 12 and Figure 13.

Furthermore, there was one three-way interaction between the factors age, gender and balance group for the UQ0 attribute "I thought the service was complicated" [$p = .043$].

Questionnaire item	Interaction	*p*-value
I found the service confusing to use.	AGE*BALANCE	.005
	GENDER*BALANCE	.010
I had to concentrate hard to use the service.	GENDER*BALANCE	.000
I felt flustered when using the service.	AGE*BALANCE	.004
I thought the service was efficient.	GENDER*BALANCE	.002
I felt under stress when using the service.	AGE*BALANCE	.003
I found the service frustrating to use.	AGE*BALANCE	.010
I thought the service was complicated.	GENDER*BALANCE	.007
When I was using the service I always knew what I was expected to do.	GENDER*BALANCE	.001
I did not feel in control when using the service.	AGE*BALANCE	.015
I felt the service was easy to use.	GENDER*BALANCE	.012
I would be happy to use the service again.	AGE*BALANCE	.002
	GENDER*BALANCE	.018
I thought the service was reliable.	GENDER*BALANCE	.024
I thought the service was polite.	GENDER*BALANCE	.026

Table 13. Summary of significant interactions for UQ0 usability attributes, all participants (*N* = 168).

4.5.5 Changes in usability ratings following SID (UQ1-UQ0)

The second set of analyses concerned the impact of the presence of a SID on participants' attitudes towards service usability. This change in attitude was computed by subtracting the UQ0 scores from the UQ1 scores; this measure – the *differential score* – is more reliable than looking at responses to the second questionnaire alone since it controls for individual differences between participants. The new values were then categorised into 'experienced a SID' (*N* = 114) and 'no SID control group' (*N* = 54), and a univariate ANOVA was then run with age, gender, balance and proposal (two levels: absence/presence, referred to as SID_Y/N in Table 14) as between-group variables.

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Between-participant effects					
AGE	.002	2	.001	.003	.997
GENDER	.198	1	.198	.585	.446
BALANCE	.253	1	.253	.749	.388
SID_Y/N	.689	1	.689	2.034	.156
AGE * GENDER	.687	2	.343	1.014	.365
AGE * BALANCE	.118	2	.059	.175	.840
GENDER * BALANCE	1.646	1	1.646	4.862	**.029**
AGE * GENDER * BALANCE	.127	2	.063	.187	.829
AGE * SID_Y/N	.601	2	.300	.887	.414
GENDER * SID_Y/N	1.338	1	1.338	3.951	**.049**
AGE * GENDER * SID_Y/N	.377	2	.188	.556	.575
BALANCE * SID_Y/N	.855	1	.855	2.526	.114
AGE * BALANCE * SID_Y/N	1.122	2	.561	1.658	.194
GENDER * BALANCE * SID_Y/N	.065	1	.065	.191	.662
AGE * GENDER * BALANCE * SID_Y/N	1.182	2	.591	1.746	.178
Error	48.743	144	.338		

Table 14. Univariate ANOVA on differential usability mean scores (UQ1-UQ0), all participants (*N* = 168).

There were no statistically significant differences for main effects of age, gender, balance or absence/presence of SID. Significant interactions were found for factors gender and balance (Figure 14) and gender and SID presence (Figure 15).

A one-way ANOVA on the new combined gender and balance independent variable (with four levels) showed overall statistical significance [$df = 3$, $F = 3.685$, $p = .014$]; Post Hoc tests revealed that the overall mean for males in the high balance group was significantly different from the males in the low balance group [$p = .008$].

Similarly, a one-way ANOVA was computed with SID presence and gender combined as the new independent variable (four levels) but showed no statistically significant differences overall [$df = 3$, $F = 2.322$, $p = .077$]. Post Hoc tests revealed that the difference between the two SID presence/absence groups in the female cohort was significantly different [$p = .047$].

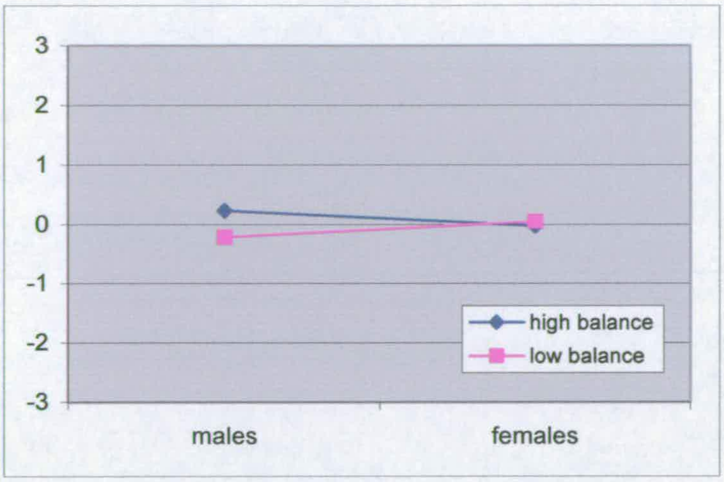


Figure 14. Overall mean usability score differences (UQ1-UQ0) grouped according to balance level and gender, all participants (*N* = 168).

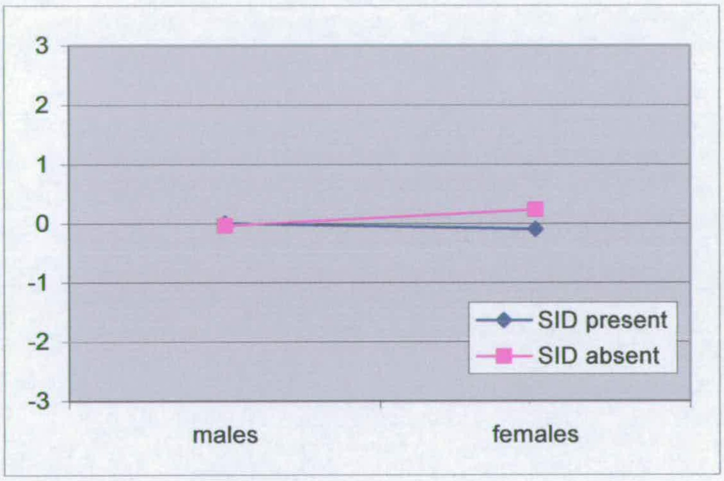


Figure 15. Overall mean usability score differences (UQ1-UQ0) grouped according to presence of SID and gender, all participants (*N* = 168).

Differential attitude scores (UQ1-UQ0) were also computed for each of the individual questionnaire statements and univariate ANOVAs (with the same between-group variables) were performed. Differential score profiles for factors age, gender, balance and SID presence are shown in Chart 4, Chart 5, Chart 6 and Chart 7 respectively. There were moderate significant main effects for age (Chart 4) with regards to *knowing what to do* when using the service [$df = 2$, $F = 3.709$, $p = .027$]; however, Post Hoc tests revealed no further significances between the three age groups. There were no significant differences for main effect of gender (Chart 5). Chart 6 shows differential scores split according to balance level.

Overall, for participants in the higher balance group, the change in attitude towards the service was slightly more positive compared to the lower balance group (Chart 6). However, only one of the attributes was statistically significant: participants in the lower balance group felt more *flustered* than participants in the high balance group [$df = 1$, $F = 5.454$, $p = .021$].

Similarly, in the SID presence/absence comparison, the change in attitude for participants in the control group condition was slightly more positive compared to participants who experienced a SID. Only two attributes were statistically significant: participants in the control group felt the *voice was clearer* [$df = 1$, $F = 5.628$, $p = .019$] and found the service more *easy to use* [$df = 1$, $F = 6.203$, $p = .014$].

The significance cut-off point used in the analyses in the current research is $p = 0.05$. When 20 such tests are performed in an ANOVA there is a 1 in 20 chance that the test will yield a false positive. Therefore, results where one or two questionnaire items are statistically significant (especially at the lower level of significance) should be interpreted with care. For example, reasons for the fact that participants in the control group found that the voice was clearer may be indicative of problems with the wording of the proposal; alternatively being unprepared for the proposal deliver might have left participants feeling that they had missed out on information. The potential for this questionnaire item to be statistically significant due to a 'false positive' should however be taken into account in the interpretation.

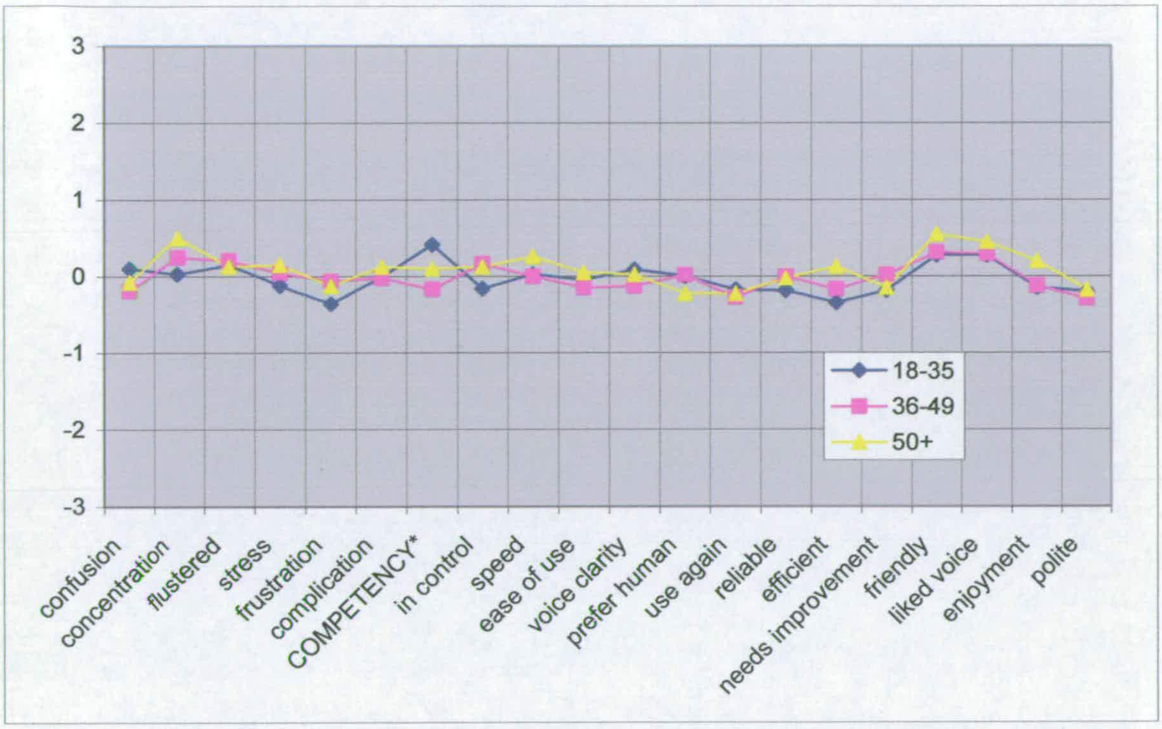


Chart 4. Differential score profiles (UQ1-UQ0), split according to age group [p*<.05], all participants (*N* = 168).**

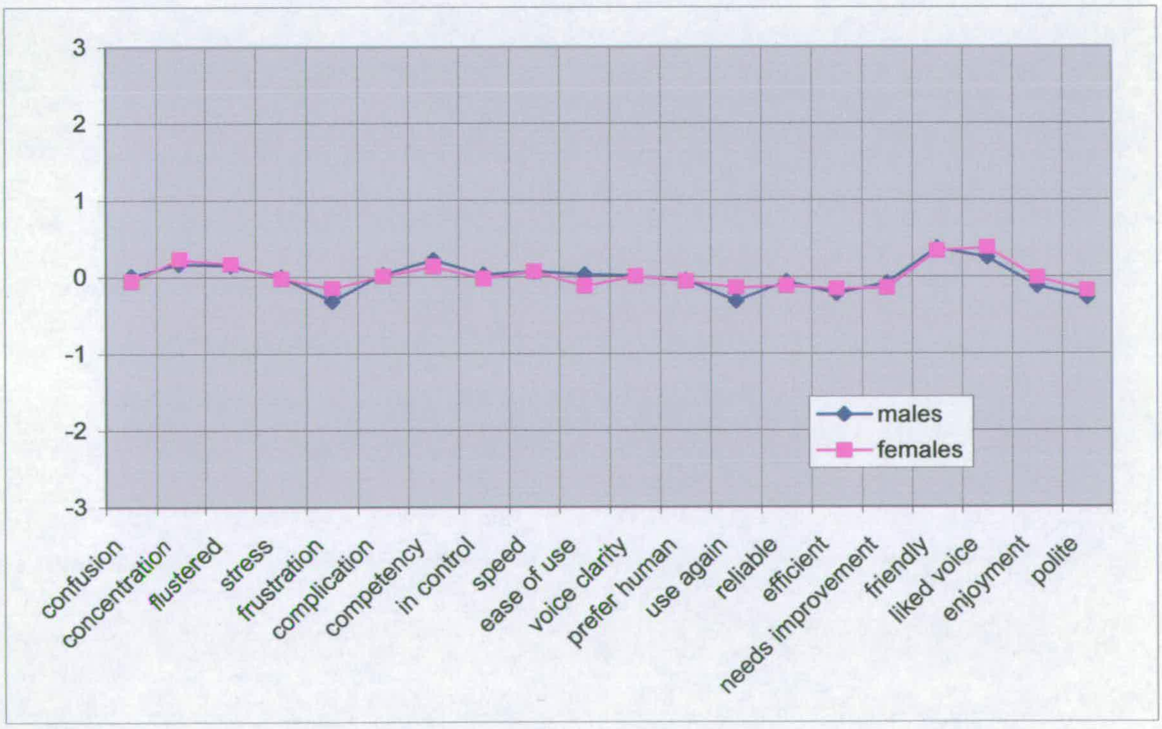


Chart 5. Differential score profiles (UQ1-UQ0), split according to gender, all participants (*N* = 168).

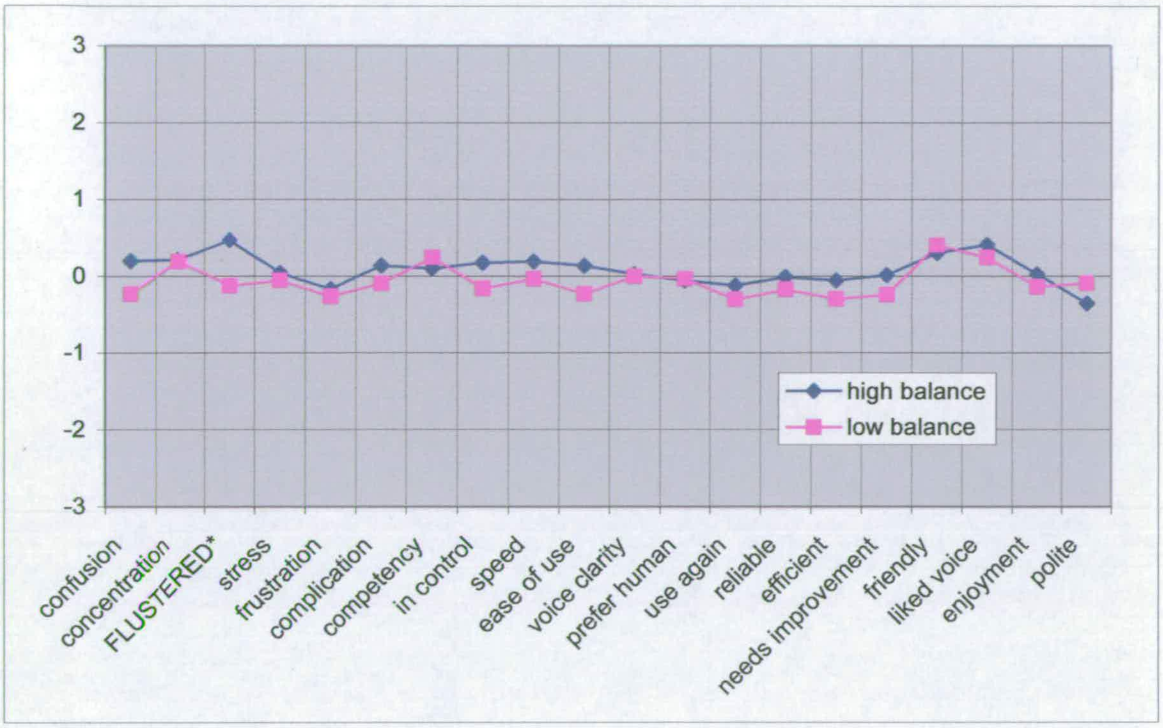


Chart 6. Differential score profiles (UQ1-UQ0), split according to balance level [*p<.05], all participants (N = 168).

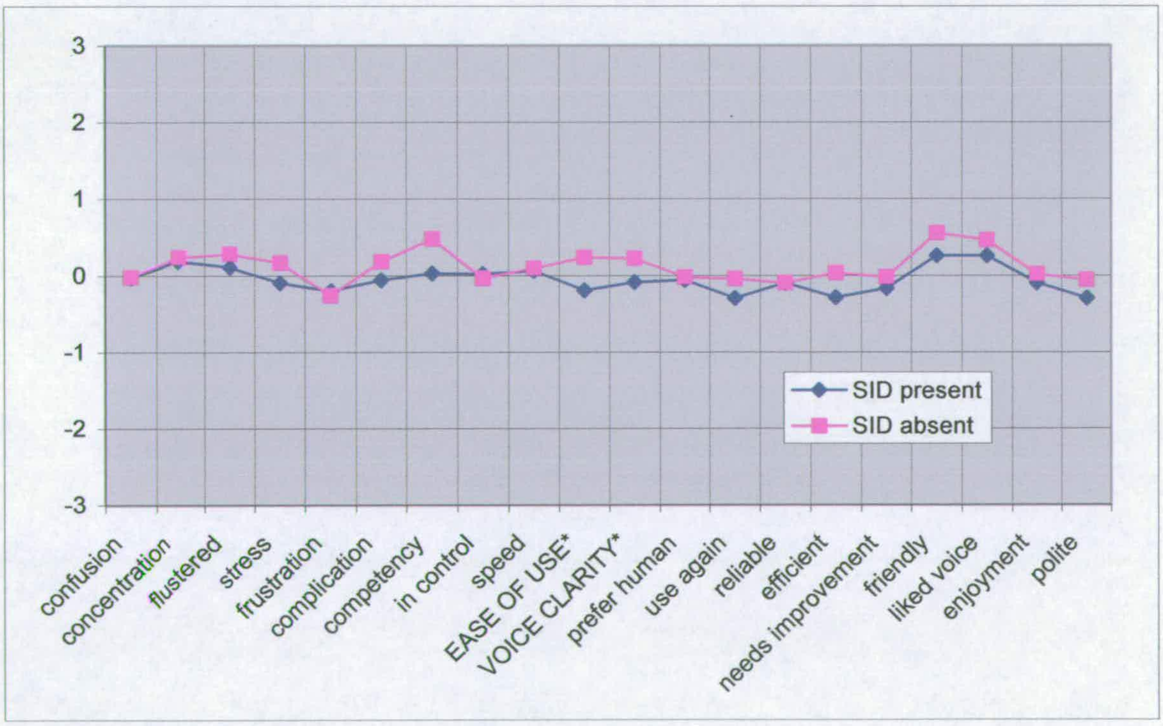


Chart 7. Differential score profiles (UQ1-UQ0), split according to presence of a SID [*p<.05], all participants (N = 168).

A number of interactions between factors were also present in the analysis. These are summarised in Table 15. Furthermore, there was one three-way interaction between the factors age, gender and SID presence for the questionnaire attribute "When I was using the service I always knew what I was expected to do" [*p* = .005].

Questionnaire item	Interaction	*p*-value
I found the service confusing to use.	GENDER*SID_Y/N	.018
I felt flustered when using the service.	GENDER*SID_Y/N	.007
I thought the service was efficient.	GENDER*BALANCE	.002
I felt under stress when using the service.	GENDER*SID_Y/N	.012
I found the service frustrating to use.	GENDER*SID_Y/N	.042
I thought the service was complicated.	GENDER*BALANCE	.010
When I was using the service I always knew what I was expected to do.	AGE*GENDER	.049
I felt the service was easy to use.	GENDER*BALANCE	.008

Table 15. Summary of significant interactions for differential scores (UQ1-UQ0), for all participants (*N* = 168).

Finally, to explore the difference in change of attitudes between the three SID strategy groups, a univariate ANOVA was run on the differential scores from participants (*N* = 114) who had experienced a SID proposal during their use of the service. Between-participant variables were age, gender, balance and SID strategy (two levels: signpost/follow-on). The results on the overall differential scores are shown in Table 16 below. There were no statistically significant differences for either main effects (although, at *p* = .057, the effect of balance level approached significance, participants in the high balance group taking a more positive attitude overall to the service) or for interaction between variables. The *Intercept* value in Table 16 represents the overall change in attitude (positive or negative) from the value 0 which represents no change in attitude. The result obtain in the analysis indicated that the presence of a SID offer did not have an impact on participant attitudes [*p* = .996].

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Between-participant effects					
Intercept	9.908E-06	1	9.908E-06	.000	.996
AGE	.552	2	.276	.652	.524
GENDER	.337	1	.337	.794	.375
BALANCE	1.578	1	1.578	3.722	.057
SID_STRATEGY	.307	1	.307	.724	.397
AGE * GENDER	.156	2	.078	.185	.832
AGE * BALANCE	.411	2	.206	.485	.617
GENDER * BALANCE	.750	1	.750	1.770	.187
AGE * GENDER * BALANCE	1.034	2	.517	1.219	.300
AGE * SID_STRATEGY	.292	2	.146	.344	.710
GENDER * SID_STRATEGY	.039	1	.039	.091	.763
AGE * GENDER * SID_STRATEGY	.275	2	.137	.324	.724
BALANCE * SID_STRATEGY	.008	1	.008	.018	.893
AGE * BALANCE * SID_STRATEGY	.238	2	.119	.280	.756
GENDER * BALANCE * SID_STRATEGY	.078	1	.078	.184	.669
AGE * GENDER * BALANCE * SID_STRATEGY	.240	2	.120	.283	.754
Error	38.156	90	.424		

Table 16. Univariate ANOVA on overall usability mean scores (UQ1-UQ0), for participants who experienced a SID offer ($N = 114$).

Univariate ANOVAs were also run on the differential scores for individual questionnaire attributes. No significant differences were found between the levels for age (Chart 8), gender (Chart 9) and SID strategy (Chart 11). Three questionnaire attributes were significant in terms of the change in attitude for the balance level comparison (Chart 10). The participants who experienced the decreasing balance rated the service as *more confusing to use* [$df = 1$, $F = 4.716$, $p = .033$], felt more *flustered* [$df = 1$, $F = 4.268$, $p = .042$] and found the service less *easy to use* [$df = 1$, $F = 5.973$, $p = .016$].

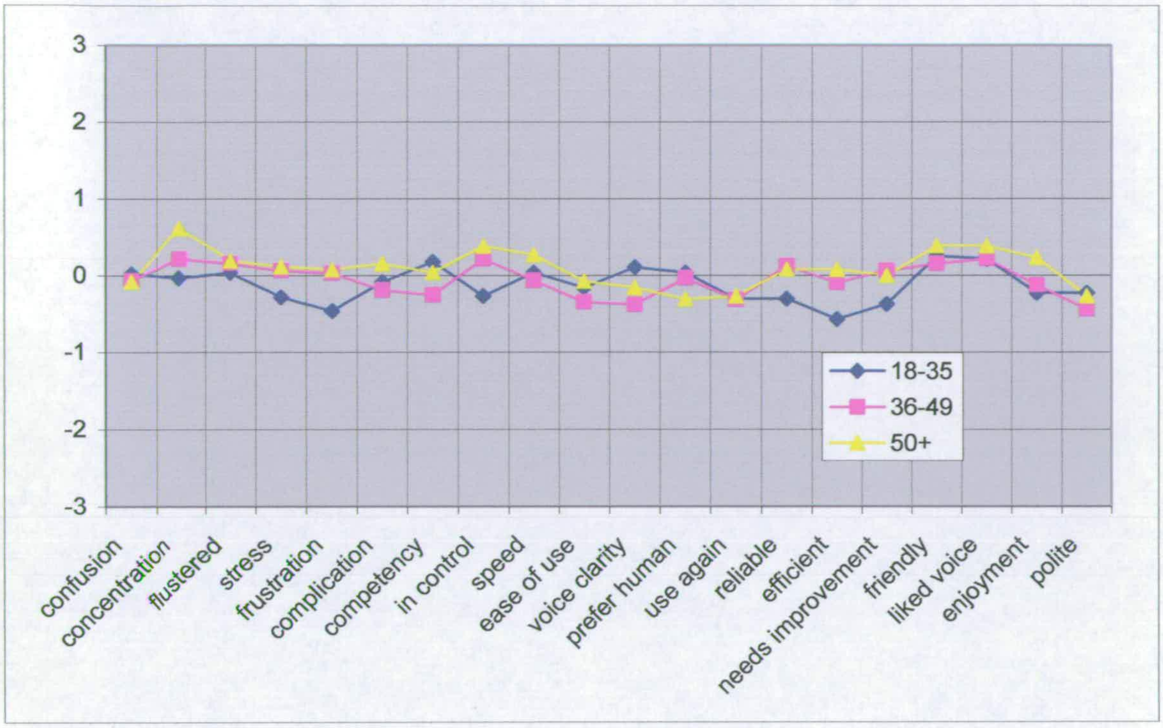


Chart 8. Differential score profiles (UQ1-UQ0), split according to age, for participants who experienced a SID offer ($N = 114$).

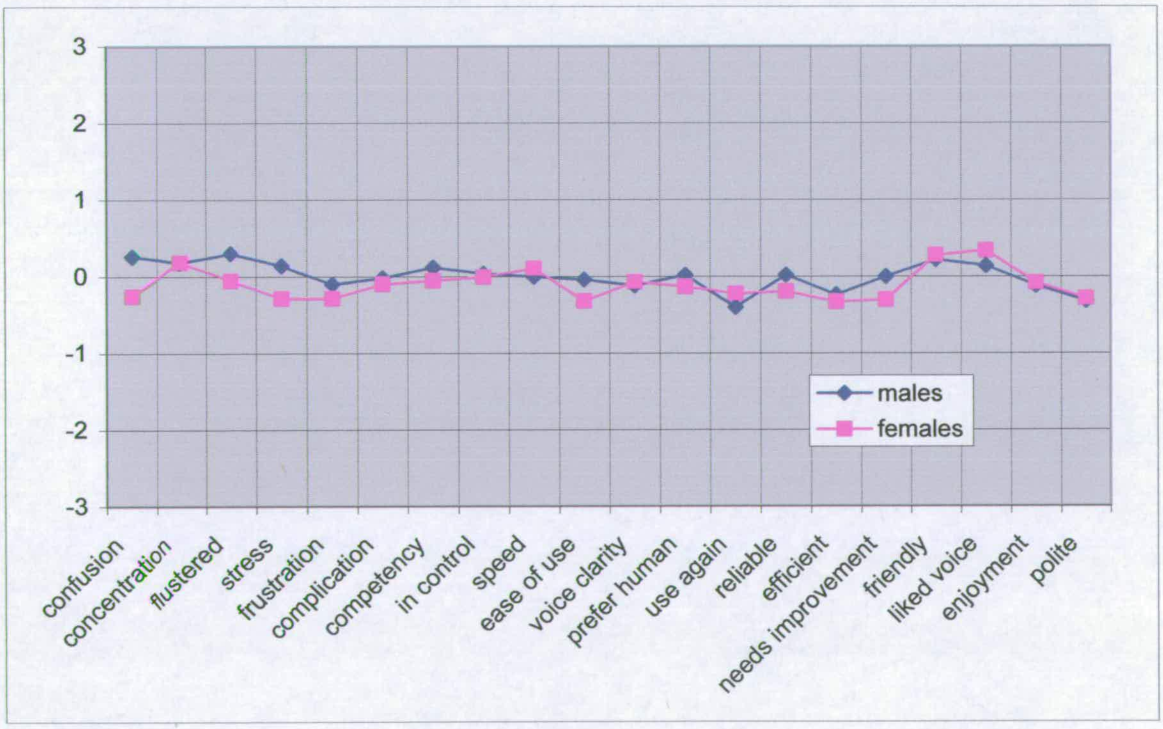


Chart 9. Differential score profiles (UQ1-UQ0), split according to gender, for participants who experienced a SID offer ($N = 114$).

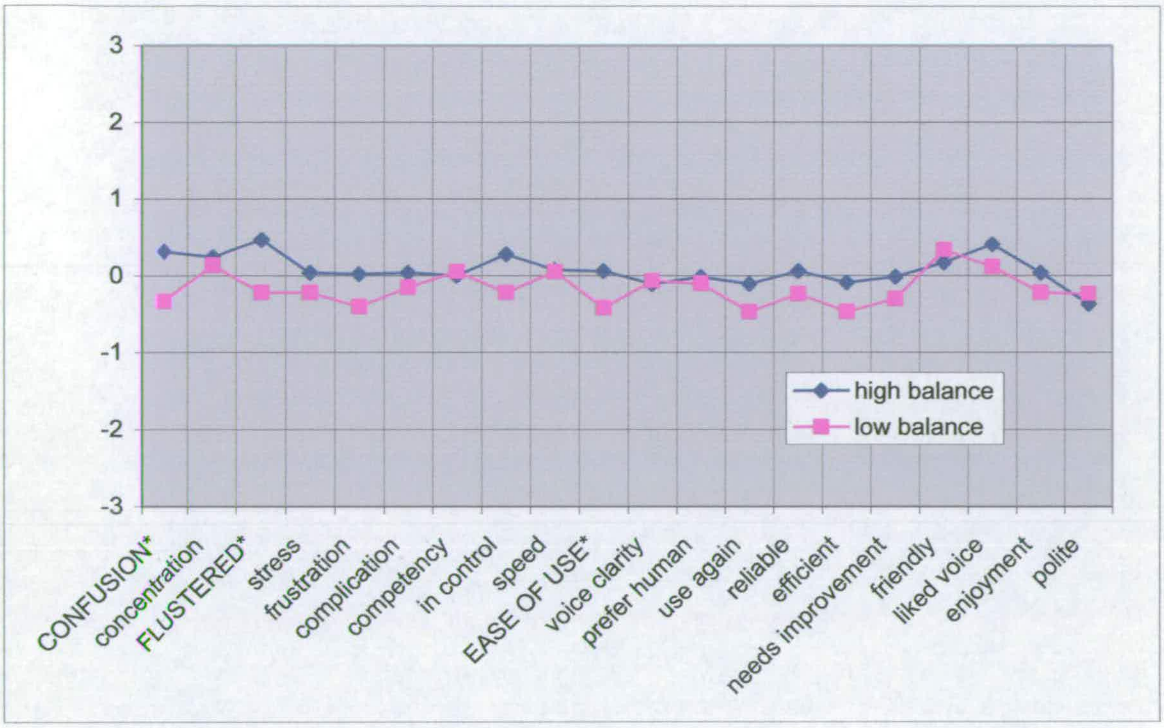


Chart 10. Differential score profiles (UQ1-UQ0), [*$p < .05$], split according to balance level, for participants who experienced a SID offer ($N = 114$).

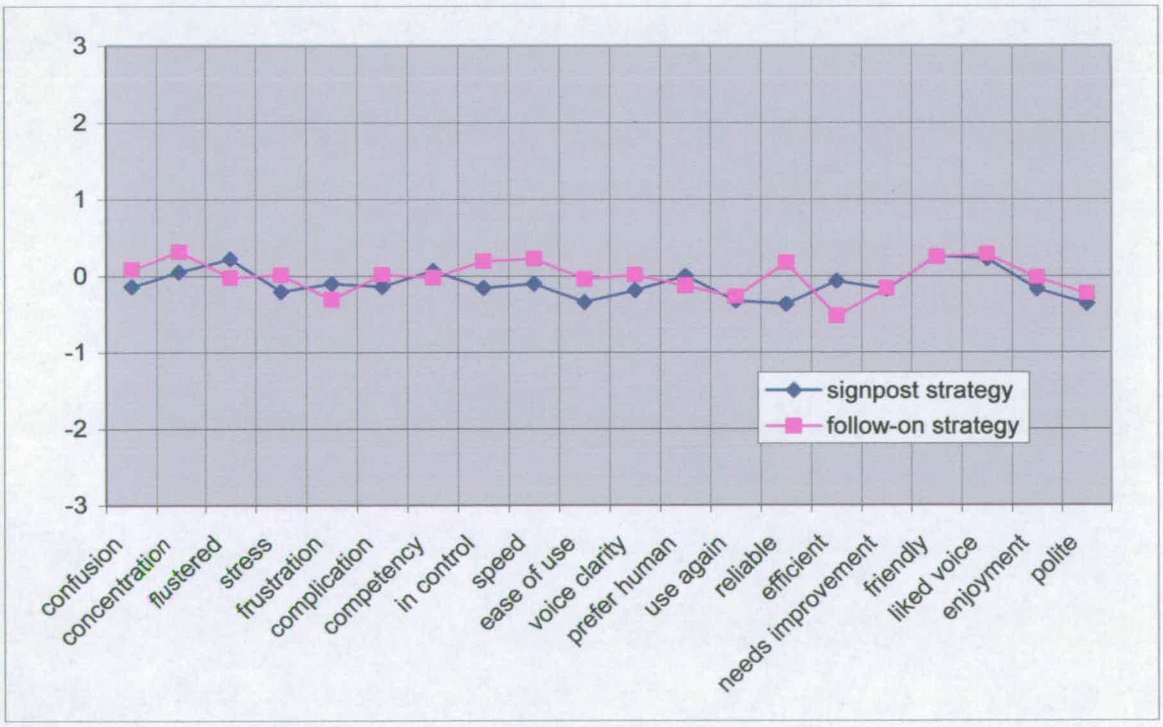


Chart 11. Differential score profiles (UQ1-UQ0), split according to SID strategy, for participants who experienced a SID offer ($N = 114$).

For each ANOVA on an individual questionnaire item, an Intercept value is returned representing the overall differential score (change in attitude) compare to value 0 (no change) for the SID proposal group as a whole. The actual differential score values are represented by the line marked as 'SID present' in the profile Chart 7 above. Five Intercept values showed significance[23]: after experiencing a proposal participants were *less happy about using the service again* [$df = 1$, $F = 6.034$, $p = .016$] and thought the service was *less polite* [$df = 1$, $F = 15.327$, $p = .000$]. However, these participants *liked the voice more* [$df = 1$, $F = 11.329$, $p = .001$]; *thought the service was more friendly* [$df = 1$, $F = 8.442$, $p = .005$]; and felt they *had to concentrate less* when using the service [$df = 1$, $F = 4.933$, $p = .029$].

There were some significant interactions between factors, as shown in the summary in Table 17. There was also one significant three-way interaction between age, gender and SID strategy in terms of the perceived *service friendliness* [$p = .027$].

Questionnaire item	Interaction	*p*-value
The service was too fast for me.	BALANCE*SID_STRATEGY	.023
I felt that the service was reliable.	AGE*SID_STRATEGY	.043
The service was friendly.	GENDER*BALANCE	.048
	AGE*SID_STRATEGY	.033
	GENDER*SID_STRATEGY	.003
I enjoyed using the service.	GENDER*SID_STRATEGY	.035

Table 17. Summary of significant interactions for differential scores (UQ1-UQ0), for participants who had experienced a SID offer (*N* = 114).

[23] A significant *p*-value would indicate that the presence of a SID had an impact on participants' attitudes towards service usability. However, such a result (particularly if a significantly positive change in attitude occurred) could be due to the habituation effect rather than SID presence (i.e. a more positive attitude towards the service as experience of the system functionality increases) and should be interpreted with caution.

4.5.6 De-briefing interview feedback

All participants took part in a structured interview after their experience with the automated service. This section concerns responses from participants who experienced the SID version of the service (N = 114).

When asked if they had noticed something different in the last phone call, all but one participant answered "yes". When asked to describe what was different, the overdraft offer was mentioned first by 71% of participants. All participants (although some only after having been prompted by the researcher) said that they had noticed the overdraft offer. The majority of participants (92%) thought it was easy to understand the information given in the overdraft proposal. There were no significant differences in participant responses between the Follow-on and Signpost SID groups.

Participants were asked why they thought they were offered an overdraft and why they had not pursued the offer. Some 37% of participants believed that the overdraft offer was made because the Bank wanted to make money whilst 72% of those with a declining balance believed this was what had triggered the offer. The main reasons for not pursuing the overdraft offer given by participants were that it was not part of the experiment tasks (29%) and that it was not actually perceived as being required (20%). Some 13% of participants replied that they were against overdrafts in principle. Only 8% felt that this was an inappropriate channel for pursuing an overdraft.

Participants were also asked to contemplate how they would have reacted to the overdraft offer "if you were using this as a *real* service". Some 40% of participant replied that they would prefer not to be offered an overdraft[24]; 31% within the Signpost SID strategy group compared to 50% within the Follow-on strategy [p = .052]. On the question whether the overdraft offer, in real life, would have discouraged them from using the service in the future, 25% of participants answered

[24] A further four participants, two in the Signpost group and two in the Follow-on group, answered that they were "not sure" and their data have been excluded from the analysis.

that they felt it would[25]; 20% of these participants had experienced the Signpost strategy and 31% the Follow-on strategy [$p = .195$].

In terms of the impact of balance level at the time of the SID offer, some 52% of participants in the high balance group felt that they would rather not have received the overdraft offer, compared to 31% in the low balance group [$p = .031$]. Some 35% in the high balance group said that the offer would discourage them from using the service in the future compared to 16% in the low balance group [$p = .028$].

4.6 Conclusions

The research in this chapter centred around three themes:

1) presence/absence of SID,

2) contrasting SID strategies and

3) impact of perceived need of an overdraft offer.

4.6.1 The impact of presence of SID

In this experiment, the predicted outcome of deploying a SID in the dialogue of an automated banking service was that it would have a negative impact on user attitudes towards the usability of the service. In fact, examining the differential mean scores for participants who experienced a SID (when compared with the no-SID control group), very few effects were found which could be attributed to the overdraft offer; there was no significant main effect of SID presence [$p = .156$] and only two individual statements in the questionnaire (*ease of use* and *voice clarity*) showed statistical significant changes. It may be concluded from these results that there is no evidence yet to suggest that making a SID offer (unexpectedly) impacts negatively on users' attitudes to the automated banking service.

[25] Here, two participants (one in each SID strategy group) answered "don't know" and their data were subsequently excluded.

This finding is reassuring, as a negative impact of SID delivery on user attitudes to the automated service is undesirable and could even prove to be counterproductive. However, participant comments from the exit interview suggest that users can be sensitive to the idea of being ‘sold’ products or services from their bank. When asked to reflect upon – in real use of the service – whether they would prefer never to be offered the overdraft, 40% of participants responded “yes” and 25% claimed that they would feel discouraged from using the service in the future. This highlights the importance of designing SIDs that deliver enough information to be useful to interested customers, at the same time as avoiding unnecessary disruption of the dialogue for customers who would rather not hear the information. It also strengthens the argument that SIDs should only occur occasionally in order to make such interruptions tolerable to customers.

In sum, contrary to the experiment prediction 1, the delivery of a SID did not significantly impact participant attitudes to the service.

4.6.2 The impact of SID strategy

The predicted effect of employing two contrasting SID strategies was that the Follow-on strategy would have a more negative impact on user attitudes towards the service. Results showed that this was not the case; there was no main effect for SID strategy [p = .397] and no individual statements showed any statistical significance. It is worth recalling that, before completing the questionnaire, participants were instructed to “think about the service in the phone call that you just made”; no reference to the overdraft offer was made until the exit interview at the end of the experiment session.

This finding suggests that no tangible differences between the two offer strategies exist (measured in terms of attitude to service usability) and that the dialogue engineer can be guided by a different set of selection criteria, such as those outlined in a requirements capture. In the exit interview, however, participants who experienced the Signpost strategy were more positively disposed towards experiencing an overdraft offer in ‘real life’ (the difference approached statistical significance) and were also somewhat less likely to feel discouraged from using the service in the future, compared to the Follow-on participant group. This indicates that there are differences between the two delivery strategies. However, it is not clear from these results exactly *what* might have triggered a more positive response to the Signpost strategy. In order to explore this issue in more detail, two objectives were identified for the follow-up experiments: firstly, a complementary questionnaire would be designed with particular focus on the issues surrounding the proposal design features; and secondly, the use of contrasting SID delivery strategies would be revisited (Experiment 3, Chapter 5).

In sum, contrary to the experiment prediction 2, the follow-on proposal did not have a significantly negative impact on user attitudes to the service compared to the signpost proposal. There was some indication (exit interview) that participants who experienced the signpost strategy were more positive to receiving the overdraft offer in real use of the service.

4.6.3 The impact of 'perceived need'

Perceived need for an overdraft is likely to have a significant impact on customers' attitudes towards receiving an unsolicited SID in the automated banking service in real life. It was established in the introduction to this chapter that real life behaviour, needs and goals are difficult to recreate in a laboratory environment. A simplified approach to the issue of perceived need was simulated in the experiment by including a declining current account balance. The prediction was that participants who received this lower balance would be more positively disposed toward the overdraft offer (i.e. an interaction between proposal presence and balance level).

Results showed that balance level did have a significant impact on participants' attitudes to the service – but these findings were in conflict with the experiment predictions. Even before experiencing the proposal, the balance level was found to have an impact on participant attitudes toward the service overall; not as a main effect but as an interaction with both age and with gender. It is not surprising that the balance level had an impact on participants' attitudes; a balance higher or lower than the participant would normally experience in real life may be perceived as a 'positive surprise', or perhaps disconcerting. Perhaps more unexpected was the finding that different age groups, males and females would respond to the variation in balance level differently.

One reason for the difference in response could be that participants experience 'ceiling effects': with a set of high scores in the UQ0 data there is no way of expressing 'improved attitude' by increasing the score in UQ1. Given that there was a difference in UQ0, it could explain a difference (in opposite direction) in the differential mean scores (UQ1-UQ0). This theory is supported by the results of the analysis of the change in attitude for the interaction between gender and balance, *after* experiencing the SID; there was a significant difference in overall differential scores for males in the two balance groups, the difference being in opposite direction compared to overall responses prior to experiencing the SID.

The exit interview showed that the majority of participants in the group with the declining balance were aware of a link between the low balance and the triggering of the overdraft offer but, contrary to the experiment predictions, this did not influence the overdraft take-up rate nor did it make participants more positively disposed towards the proposal (UQ1-UQ0). The psychology of borrowing is likely to be multi-factorial: there is unlikely to be one driver which explains people's propensity to take up offers of overdrafts. The evidence from this experiment supports this view insofar as the single predictor of likelihood of taking up the overdraft offer (declining balance) did not in fact influence participants' behaviour or attitudes in the way expected. Further research is needed in order to establish the true impact of perceived need on user attitudes to automated banking. Such follow up studies are likely to involve real-world behaviour or complex scenario building; issues which are outside the scope of the current research. The remaining chapters in this thesis will be concerned with the dialogue engineering and usability aspects of SIDs in automated telephone banking.

In sum, contrary to experiment prediction 3, the 'perceived need' (established through a low balance) did not have a significant positive impact on user attitudes to the service with the overdraft offer. There was some indication (exit interview) that participants who experienced a decreasing balance ('perceived need') were more positive to receiving the overdraft offer in real use of the service.

Chapter 5

Observe due measure, for right timing is in all things the most important factor.

- Hesiod (~700 BC), Greek poet -

Experiment 2 – Dialogue location of system-initiated digressive proposals

5.1 Introduction

Following on from Experiment 1, which investigated strategies for *how* to offer an overdraft, the current research extends this to explore *where* to locate a SID. This is achieved by deploying Signpost-style overdraft offers in three contrasting locations within the dialogue of an automated telephone banking service. The prompt register was, as in the experiment in the preceding chapter, worded such that it toned with the existing prompts in the automated service as much as possible. SID location and prompt register form the core dialogue engineering design issues in the current experiment.

The current dialogue design features 'hidden' menu options, that is, new service options introduced in the SID offer are active – but not explicitly listed – in the main menu; this approach enables new options to be introduced into the automated banking service without increasing the length or complexity of the main menu. The future success of these SIDs will rely largely on the users' ability to successfully locate and select the hidden menu option within the automated service dialogue. In order to determine the viability of the hidden menu approach, the current experiment included the task of applying for an overdraft and participant performance data (task completion and navigational route through the dialogue) were explored.

Following on from Experiment 1, it was established that the usability evaluation tool (20-statement Likert-style questionnaire) could benefit from being complemented with an additional set of attributes which are aimed at eliciting participants' reactions towards the SID; thus extending the evaluation of service usability to encompass attitude towards the SID component itself. The construction of this additional 'SID questionnaire' will be included in the Section 5.4.4 in this chapter.

5.2 Design objectives

Three design objectives have been introduced and will be explored in the current experiment: delivery location, prompt register and offer completion. The rest of this section is concerned with a more detailed description of the design consideration and the implementation particulars involved.

5.2.1 Dialogue engineering objective 1: Location for delivery

The SID prompt and dialogue style employed in this experiment are based on the Signpost strategy of overdraft offers developed in Experiment 1. The main characteristics of the Signpost offer are a short message, embedded within the normal dialogue call flow, informing customers about the availability of the overdraft facility and how to request an overdraft within the automated banking service (in this case by saying "overdraft" at the menu of services). The Signpost strategy was selected for this experiment as it lends itself to be used in all three of the chosen locations of the dialogue: Welcome, ID&V (Identification and Verification) and Transaction. The locations of these three offers are marked in the dialogue flow-chart, Figure 16.

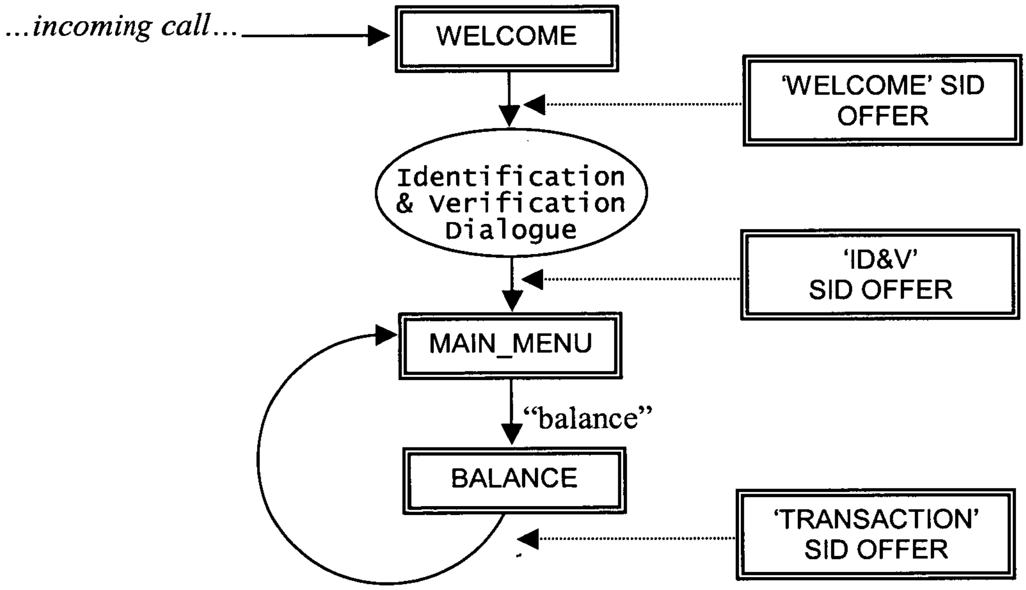


Figure 16. Overdraft offer dialogue, marking the locations of the Welcome, ID&V and Transaction offers.

The first of these three locations – the Welcome offer – follows the introductory "Welcome to PhoneBank *Express*" message in the very initial stage of the phone call to the service. The intention of locating the SID here would be to inform *all* customers about the availability of the overdraft option accessible from the main menu. This location only really lends itself to SID offers with a general nature that apply to all customers (the issue of making offers to ineligible customers is addressed in Section 5.2.3 below). Due to its location in the dialogue it might be expected that the proposal part of the Welcome message will pose a lower risk of distracting the customer from their task at hand. However, an upfront message like this every time the caller contacts the automated service may be perceived as annoying, especially if the caller does not know how to bypass (barge through) the offer.

The second proposal variant was located immediately after a caller had been successfully identified by a valid membership number and verified successfully using their secret TIN. This version is referred to here as the 'ID&V' (identification and verification) proposal. As the customer's identity has been established the offer can be targeted to suit the individual's particular banking situation; additionally, the risk of offering products to ineligible customers is reduced.

Finally, the third 'Transaction' offer was a nested prompt that followed a particular transaction or sub-dialogue in the service – in this case after a balance request, immediately after the amount had been played. The Transaction location enables references to certain account details, a particular transaction, topic or service to be made in the SID. The ability to create a logical link between offer information and specific details in the dialogue can be useful but it could also be potentially more distracting to the customer who is heavily involved with the task at hand and might miss out on important product information in the SID. There is also the risk that, having obtained the primary information (e.g. balance), the caller hangs up and does not hear the SID offer at all.

5.2.2 Dialogue engineering objective 2: Prompt register

The wordings of the SID overdraft offers were optimised based on their location in the dialogue. The Welcome location proposal was worded in such a way as to be general in nature and applicable to all callers:

> ***"[Welcome to PhoneBank Express.] We've added a new overdraft facility to this service. To find out more, just say overdraft at the menu of services".***

The ID&V location offer could – since the caller had been identified – be made customer oriented with reference to specific amounts and accounts:

> ***"You might like to know that you can have an overdraft of 400 pounds on your current account. To find out more, just say overdraft at the menu of services".***

Similarly, the Transaction location offer could also be made customer oriented and – since the balance information for the current account has just been played – an anaphora ("this" account) was used instead of the account name.

> ***"You might like to know that you can have an overdraft of 400 pounds on this account. To find out more, just say overdraft at the menu of services".***

The dialogue module for handling overdraft requests (when selected from the main menu) remains unchanged from Experiment 1 and is detailed in Appendix 2.1. The overdraft limit was set to a fixed sum of 400 pounds and the caller could either accept or reject this amount; there was no opportunity for negotiating a lower or higher amount.

5.2.3 Contributing factors: overdraft completion and applicant eligibility

Coupled with user acceptance and application usability, the future success of these system-initiated proposals will rely largely on the users' ability to successfully locate and select the hidden menu option within the automated service dialogue. The experiment design included an extra phone call in which participants were instructed (following a balance and order of a statement) to also apply for an overdraft. Participants' inputs and navigational paths through the dialogue were logged and analysed.

Furthermore, because the Welcome SID offer occurs prior to the identification of the customer, a customer's eligibility for an overdraft cannot be established at this stage in the dialogue; this may subsequently lead to situations where an applicant who pursues the offer has to be turned down. In order to investigate this effect, half of the group of participants who experienced the overdraft offer in the Welcome location were subsequently subjected to a rejection message when they went on to apply for an overdraft in their final call to the service.

5.3 Experiment predictions

The primary aim of the experiment was to assess the relative effectiveness of the three offer locations described above. The main experiment predictions were as follows:

1. Based on the findings from Experiment 1, it was predicted that the delivery of a SID (overdraft offer) would have little or no impact on participants' attitudes towards the usability of the automated service.

2. The three SID proposals were similar in duration and the style of register applied. Therefore, it was predicted that SID location would have little impact on participants' attitudes towards the automated service. It was predicted that participants who experienced the Welcome SID might react negatively to the fact that the offer would be experienced multiple times with repeat use of the automated service. Furthermore, participants who experienced the Transaction SID location might find the offer more intrusive and interruptive compared to the Welcome and ID&V SID groups.

3. The potential problem with the Welcome SID location is having to turn down ineligible applicants; this is predicted to have a negative impact on attitudes towards the automated service.

5.4 Method

Chapter 3 provided an overview of the experiment method adopted in the current research. This section provides further details that are relevant and specific to SID Experiment 2.

5.4.1 Design

As indicated above, three different SID locations were explored. Experiment analyses rely mainly on a between-group design; repeated-measures, within-group comparisons were indirectly achieved by running between-group analyses on the change in attitude – the differential scores – computed by subtracting baseline questionnaire scores (obtained after two practice phone calls) from questionnaire scores obtained after the third (SID offer) and fourth (overdraft request) phone calls. Grouping (between-group) variables included age, gender, offer location (Welcome, ID&V and Transaction) and outcome of overdraft request (rejected or accepted).

5.4.2 Participants

A cohort of 114 participants (50 males and 64 females) contributed to the evaluation in Experiment 2[26]. Participants were recruited from the general public and only five of them had previous experience of using PhoneBank *Express*.

5.4.3 Materials

Participants were given the following personal banking details (described further in section 3.3): a membership number, a TIN, details of two accounts (one savings account and one current account). In contrast to Experiment 1, participants were not given any priming materials or pamphlets on how to operate the automated service; instead they were told that they would be using an automated telephone banking service – PhoneBank *Express* – and they could speak their commands or use the

[26] Data sets from 10 participants have been excluded from the analysis as they did not complete all phone calls to the automated service, or because they had not experienced the SID offer. A further five participant data sets also had to be excluded; these included three data sets where participants never had the opportunity to try the overdraft task in call four (had hung up or were transferred), and two data sets had corrupt system log files which could not be used.

buttons on the telephone keypad. No other instructions on how to use the service (i.e. which buttons to press or which words to use) were given. The sentence in the third error-level prompt (marked 'MEM_NUM', error=2 in section 3.3.2) "the membership number as printed on your membership card" was removed to accommodate the change in priming.

The main reasons for removing the priming from the experiment process was to avoid inadvertently introducing an extra variable that is difficult to control for. In Experiment 1 participants were given a few minutes to read through the PhoneBank *Express* material at their own leisure (no obligations on how much or what to read). Some participant read the material carefully while other participants just read specific sections of the information. In light of this finding, participants were not given any priming and the cheque task (which may require obtaining some more detailed understanding of the 'item search' service option) was changed to 'order statement'. Subsequently, the system prompts was considered to give sufficient and adequate guidance to users on how to use the service.

5.4.4 SID questionnaire

An additional questionnaire was constructed with the aim of capturing participants' attitudes toward the SID dialogue (this will hereon be referred to as the 'SIDQ' questionnaire). The final set of 16 questionnaire items centred around a number of 'themes' relating to the characteristics of interrupting the flow of the dialogue and making an unsolicited offer (the complete questionnaire as it appeared in the experiment is included in Appendix 1.5). The SIDQ items were constructed following the 'Likert' format described in 2.6.2, with an equal balance of positive and negative statements. The first questionnaire construct concerns the 'social aspects' of introducing a SID into the conversation and the impact that this may have on the caller's perception of the overdraft proposal:

The overdraft proposal was **polite**.
I found the overdraft proposal **intrusive**.
The overdraft proposal was **annoying**.
The overdraft proposal **distracted me** from what I was trying to do.

Secondly, a set of four questionnaire items were devised to address 'channel suitability' – the degree to which the SID offer is perceived to fit in with the rest of system dialogue and the appropriateness of using an offer as a method for introducing this kind of information.

The overdraft proposal was **too long**.
The overdraft proposal **interrupted the call** too much.
It is **appropriate** to have overdraft proposals in this kind of automated service.
The overdraft proposal is an **efficient method** for giving product information.

The third construct related to the caller's trust in the service: if the caller considers the automated service a reliable source for receiving this kind of offer information, or if they would prefer a human agent to approach them about the overdraft instead. Furthermore, the construct aimed to establish whether or not callers found that they could carry out the overdraft application through the automated service.

I would **trust** the automated service to give me appropriate overdraft information.
I'd **prefer** an overdraft proposal to be made by a **human agent** rather than the automated service.
If I needed an overdraft, I would be **happy to apply through the automated service**.
I **wouldn't rely solely on the automated service** when seeking an overdraft.

The fourth construct concerned the callers' perceived understanding of the overdraft proposal: whether or not the information was easy to understand and whether or not callers found that they had received enough information to be able to apply for an overdraft through the automated service.

I now **know how to use** the automated service to apply for an overdraft.
The overdraft proposal was **easy to understand**.

Finally, it was desirable to establish whether or not callers appreciated the overdraft proposal (its perceived helpfulness) or felt that they were interested in the information.

The overdraft proposal information was **helpful.**

The overdraft proposal **was irrelevant to me**.

The questionnaire statements were randomly entered into the questionnaire, followed by a row of seven tick boxes where respondents could express degrees of 'strongly agree' to 'strongly disagree'. Before administering the SIDQ questionnaire the researcher took care to establish that the participant had noticed the overdraft proposal and point out that the questionnaire related to the participant's attitudes to the overdraft proposal.

5.4.5 Procedure

Participants were assigned to one of the three experiment conditions at random: Welcome SID location (50% in the *eligible* for overdraft group and 50% in the *ineligible* for overdraft); ID&V SID location; and Transaction SID location. Upon arrival, the participant was greeted and asked to take a seat by the telephone. A further modification to the experiment procedure was to move the demographic and technographic information capture (questionnaire available in Appendix 2.3) to the very end of the experiment session; this was done in order to reduce the risk of inadvertently making participants commit themselves to a specific 'profile' that they feel should be upheld and which, as a consequence, may have an impact on their ensuing responses.

The participant received instructions about the experiment and was then given a sheet containing the fictitious persona details (see Section 3.7 for a more detailed description). The participants' tasks were to make telephone calls (four in total) to the automated banking service to find out the balance of 'their' current account and then order a printed statement[27] for 'their' savings account. Participants were asked to take a note of the balance (the balance fluctuated slightly in order to maintain

[27] The 'order statement' task was chosen over the 'find cheque' task used in Experiment 1 as participants were given no priming on the functionality of the automated service; thus, main menu options that are not particularly self-explanatory, such as the 'item search', are likely to be confusing and it might not be clear that cheque information is actually available through selecting 'recent transactions'.

participants' engagement in monitoring the accounts). These two tasks were then repeated through each of the four phone calls. On the third call participants were exposed to an overdraft proposal and in the fourth call they were instructed (in addition to the balance and order statement tasks) to apply for an overdraft and to write down the result of the application. The task sheet for the fourth phone call is available in Appendix 3.1. The dialogue functionalities for the balance and order statement tasks are explained in further detail in sections 3.3.5 and 3.3.6.

The experiment session proceeded in a number of clearly defined stages which are outlined in Table 18 below. All participants in the experiment made their first two phone calls to the same, core, version of the automated service (without SID offers). Following the completion of the second call, participants were then asked to complete an attitude questionnaire (Appendix 1.4) to establish the reference level of the usability of the service; this questionnaire will be referred to here as 'UQ0'.

Experiment stage	Experiment condition	Materials used
Welcome, introduction, priming	Same for all participants	Persona details
Two phone calls to core service	Same for all participants	Task sheets (in each call obtain balance; order printed statement)
Usability assessment	Same for all participants	Usability questionnaire (UQ0)
One phone call to service with proposal **3 versions implemented**	1: Welcome proposal location	Task sheet (obtain balance; order printed statement)
	2: ID&V proposal location	
	3: Transaction proposal location	
Usability assessment		Usability questionnaire (UQ1) SID offer questionnaire (SIDQ)
One phone call to service **2 versions implemented**	1: Overdraft request accepted (all participants, except 50% of the Welcome proposal group)	Task sheet (obtain balance; order printed statement; apply for an overdraft)
	2: Overdraft request rejected (50% of the Welcome proposal group)	
Usability assessment		Usability questionnaire (UQ2)
De-briefing interview		De-briefing interview Demographic questionnaire

Table 18. Overview of Experiment 2 procedure.

During the third call to the service, participants experienced the overdraft proposal. After this phone call participants completed the same attitude questionnaire but focussing on their experience of the service in last call (referred to as 'UQ1'). Additionally, after establishing that the participant had noticed the information about the overdraft[28], the 'SIDQ' was administered with the instructions that the use of 'overdraft proposal' in the questionnaire referred to the overdraft information experienced.

In the fourth and final call to the automated service, participants were instructed to (in addition to the balance and order statement task) also apply for an overdraft and note down the result of their request. Following the completion of this phone call, participants completed another usability questionnaire ('UQ2'). The session was then ended with a de-briefing interview and the demographics/technographics questionnaire. A summary of the experiment design is provided in Table 19.

[28] Participants were asked if they "had notice something different with the service in the last call". Participants who did not mention the overdraft message themselves were prompted for this information directly. All participants (in the SID delivery group) had heard the overdraft offer.

Title	**Experiment 2: location for delivery**	
Design		One independent sample, between-subjects design adopted
Predictions	E2.1	The system-initiated digression would have no impact on participant attitudes to service usability.
	E2.2	The location of delivery would have little impact on participants attitude to service usability. The Transaction location would be perceived to be more intrusive.
	E2.3	Having to turn down ineligible customers would have a negative impact on attitudes toward service usability.
Independent variables	1	Application: service version (3 levels)
	2	Application: overdraft request outcome Welcome group (2 levels)
	3	Participant: gender (2 levels)
	4	Participant: age group (3 levels)
Dependent variables	1	Usability questionnaire, 'UQ0', 'UQ1' and 'UQ2' (1-7 Likert scale) SID questionnaire, 'SIDQ' (1-7 Likert scale)
Other data		De-briefing interview
Location		University Research Centre, central Edinburgh
Participant cohort		*N* = 120 (target, 30 participants in each experiment condition)
Remuneration		£20
Duration		Approximately 40 minutes

Table 19. Summary table of the SID location Experiment 2.

5.5 Results

The results analysis presented in this section was based on data entries from participants (N = 114) who had managed to successfully complete all their four phone calls to the service. Results include: demographic/technographic details, task completion rates, navigational path (for the overdraft application task), usability and SID evaluation ratings and de-briefing interview data.

5.5.1 Demographic/technographic data

Table 20 details the participant age and gender distribution for each experiment condition. The sample was overall well balanced by gender, although some cells were slightly over represented. There was also a bias evident towards the youngest age group, consequential from the recruitment process.

About 45.5% of participants (N = 50) stated that they had used an automated telephone banking service for their personal banking needs, prior to taking part in the experiment.

		Experiment condition			
Age group	Gender	Welcome proposal location	ID&V proposal location	Transaction proposal location	Total
18-35 years	Male	13	11	7	31
	Female	13	15	16	44
36-49 years	Male	2	1	4	7
	Female	2	1	2	5
50+ years	Male	6	2	4	12
	Female	3	7	5	15
Total		39	37	38	N = 114

Table 20. Analysis of participant cohort by age, gender and experiment condition.

5.5.2 Task completion

Task completion rates were based on system log data and required: (1) that the participant had taken a (correct) note of the current account balance; (2) that they had completed a request for a printed statement on their savings account; and (3), in the final phone call, that they had successfully requested an overdraft (irrespective of their request having been accepted or rejected). Overall, task completion rates for both the balance request and statement order were high (>90%, Table 21). In contrast, only 63.2% of participants requested an overdraft.

	Call 1	Call 2	Call 3	Call 4
Balance	112 (98.2%)	113 (99.1%)	113 (99.1%)	114 (100%)
Order statement	103 (90.4%)	106 (93.0%)	109 (95.6%)	110 (95.6%)
Overdraft request	N/A	N/A	N/A	72 (63.2%)

Table 21. Task completion success rates for each of the four phone calls to the automated service.

Overdraft task completion rates were further explored in terms of age (Table 22), gender (Table 23) and SID offer location (Table 24). Chi-square tests showed that the difference in overdraft task completion rates between males and females was statistically significant [p = .006].

	18-35	36-49	50+
Requested overdraft	45 (60.0%)	7 (58.3%)	20 (74.1%)
Did not request overdraft	30 (40.0%)	5 (41.7%)	7 (25.9%)

Table 22. Overdraft task completion rates (fourth phone call) based on age group factor.

	Males	Females
Requested overdraft	39 (78.0%)	33 (51.6%)
Did not request overdraft	11 (22.0%)	31 (48.4%)

Table 23. Overdraft task completion rates (fourth phone call) based on gender.

	Welcome	ID&V	Transaction
Requested overdraft	28 (71.8%)	21 (56.8%)	23 (60.5%)
Did not request overdraft	11 (28.2%)	16 (43.2%)	15 (39.5%)

Table 24. Overdraft task completion rates (fourth phone call) based on SID offer location.

In each of the four phone calls to the service, participants tended to carry out the tasks in the order prescribed on the task sheet, i.e. balance first and then order statement. Once the participants had completed the balance request they were faced with the prompt "would you like another service?". At this point in the dialogue the order statement task could be carried out in three different ways. The participant could volunteer the "order statement" command and thus bypass the menu listings altogether. Alternatively, the participant could instead answer "yes" causing the service to loop back to the Main Menu 'a' prompt (Figure 4): "please select balance, recent transactions or another service". Again, here participants could volunteer "order statement" or say "another service" in order to hear the Main Menu 'b' half of the menu listing. And finally, participants could then say "order statement" after waiting to hear the option being listed in Main Menu 'b'. Based on at which point in the dialogue "order statement" had been uttered, participants were grouped according to their propensity to 'volunteer' input information. Those who had said "order statement" before hearing it listed in the Main Menu 'b' in at least one of their four phone calls were labelled 'volunteered' (N = 39); the remaining participants were labelled 'not volunteered' (N = 75).

Table 25 shows the overdraft task completion rates, analysed in terms of participants' volunteering behaviour. The overdraft task success ratio for the 'volunteered' participants was significantly higher than for the 'not volunteered' participants who had said "order statement" only after hearing it listed in the Main Menu 'b' in all of their four calls ($N = 114, p = 0.001$)[29]. Only six participants (15.4%) in the volunteer statement group failed to complete the overdraft task (all of them having listened to both Main Menu 'a' and 'b').

	Volunteered	Not volunteered
Requested overdraft	33 (84.6%)	39 (52.0%)
Did not request overdraft	6 (15.4%)	36 (48.0%)

Table 25. Overdraft task completion rates (fourth phone call), based on "order statement" volunteering behaviour.

The findings above suggest that volunteering behaviour and gender have a significant impact on participants' ability to complete the overdraft task. The question therefore arises whether or not there are any significant variations in the propensity to volunteer input between males and females. Results showed that 42.0% (N = 21) of males and 28.1% (N = 18) of females volunteered "statement" in at least one of their phone calls, however, this difference was not strong enough to produce a statistically significant effect ($N = 114, p = 0.164$).

[29] Similar statistical findings were obtained when using volunteering behaviour from the first two phone calls only (N = 114, p = 0.002). This indicates that the system-initiated proposal had little or no impact on participants' volunteering behaviour.

Figure 17 shows the overdraft completion ratio for males and females, split according to statement volunteering behaviour. Within the male participant group, volunteering behaviour did not have a significant effect on overdraft task completion ($N = 50$, $p = 0.319$); some 86% ($N = 18$) of the male volunteered subset ($N = 21$) completed the overdraft request compared with 72% ($N = 21$) in the male not volunteered subset ($N = 29$). In contrast, volunteering behaviour in the female participant group had a strongly significant impact on overdraft task completion ($N = 64$, $p = 0.002$); some 83% ($N = 15$) of the female volunteered subset ($N = 18$) completed the overdraft request compared with 39% ($N = 18$) in the female not volunteered subset ($N = 46$).

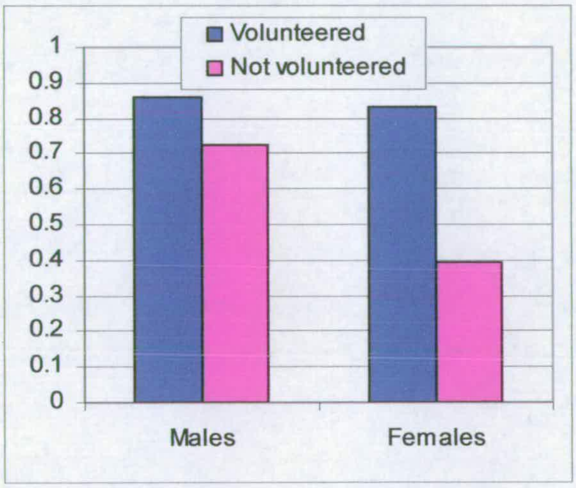


Figure 17. Overdraft task completion ratio for male and female participants, split according to whether or not they had volunteered "order statement" during at least one of their four phone calls.

As described in the section 5.4.2, about half of participants ($N = 50$) stated that they had prior experience of using an automated telephone banking system for their personal banking. Of these, 58.0% ($N = 29$) of participants with previous telephone banking exposure completed the overdraft, compared to 68.3% ($N = 41$) of participants who had no previous exposure (non-significant, $N = 110$, $p = 0.321$). In terms of volunteering behaviour, 30.0% ($N = 15$) of participants with previous exposure volunteered statement at least once during their four phone calls, compared with 36.7% ($N = 22$) of those with no previous exposure.

5.5.3 Usability ratings prior to experiencing the SID (UQ0)

In the third phone call to the service, participants experienced a system-initiated overdraft offer in one of three locations. Participants' attitudes towards the service were measured both following their second practice phone call (UQ0), immediately prior to experiencing the SID, and then after completing the phone call with the SID delivery (UQ1). Responses to the usability questionnaires were analysed, both in terms of overall mean scores and according to means for individual attributes (per statement analysis).

A univariate ANOVA was run on participant responses from UQ0 with age and gender as between-group variables (Table 26). There was a moderate effect for age (questionnaire mean scores: 18-35 *M* = 4.70; 36-49 *M* = 5.16; 50+ *M* = 4.33). Post Hoc tests revealed that the difference between the oldest (with the lowest mean score overall) and the mid-age group (with the highest mean score overall) was significant [*p* = .049].

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Between-participant effects					
AGE	6.141	2	3.071	3.070	**.050**
GENDER	.084	1	.084	.084	.773
AGE * GENDER	.408	2	.204	.204	.816
Error	108.022	108	1.000		

Table 26. ANOVA on overall usability mean scores (UQ0).

Univariate ANOVAs (with between-group variables age and gender) were also performed on each of the individual 20 UQ0 statements; score profiles for main factors are shown in Chart 12 and Chart 13. Only two individual questionnaire items were statistically significant with regards to age: the preference for *speaking to a human being* [$df = 2$, $F = 3.096$, $p = .049$] with Post Hoc tests revealing statistically significant differences [$p = .037$] between the mid-age group and the older age group; the perceived *friendliness* of the service [$df = 2$, $F = 5.016$, $p = .008$], again, with the difference between the mid- and older age groups statistically significant [$p = .007$]. There were no significant effects found for gender (Chart 13).

There were no significant interactions between factors in the analyses of UQ0.

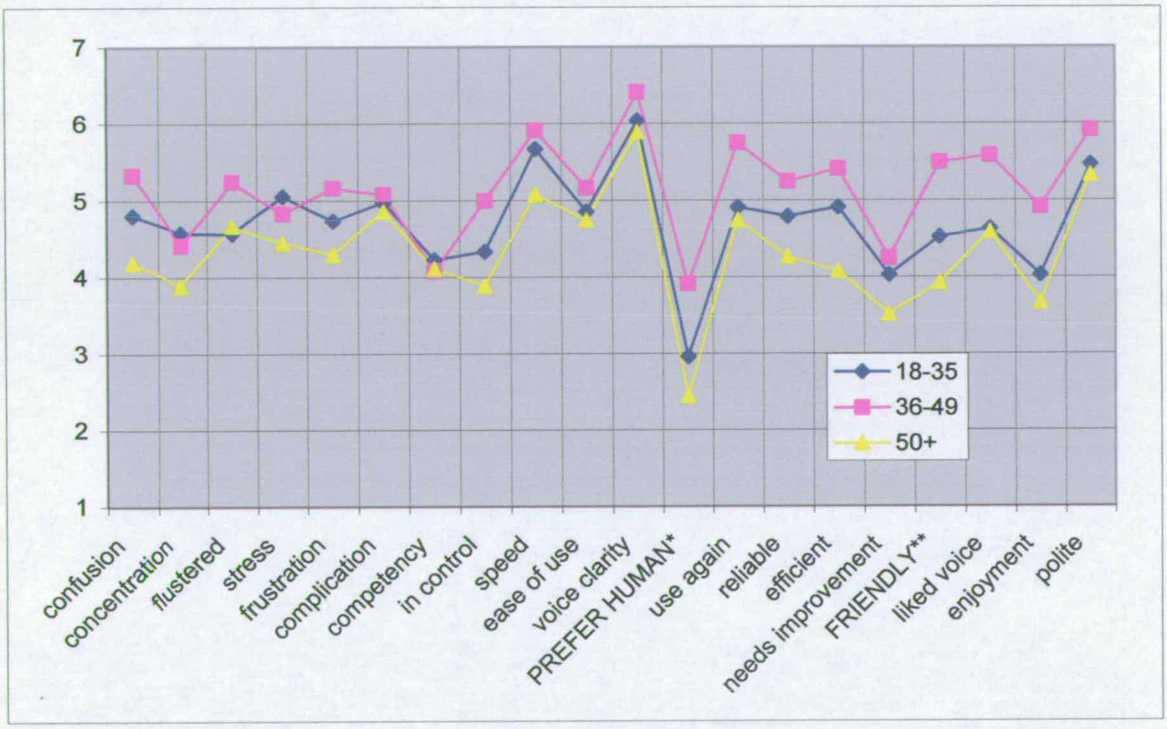


Chart 12. Main scores for UQ0 attributes, split according to age factor with three levels [*p<.05; **p<.01].

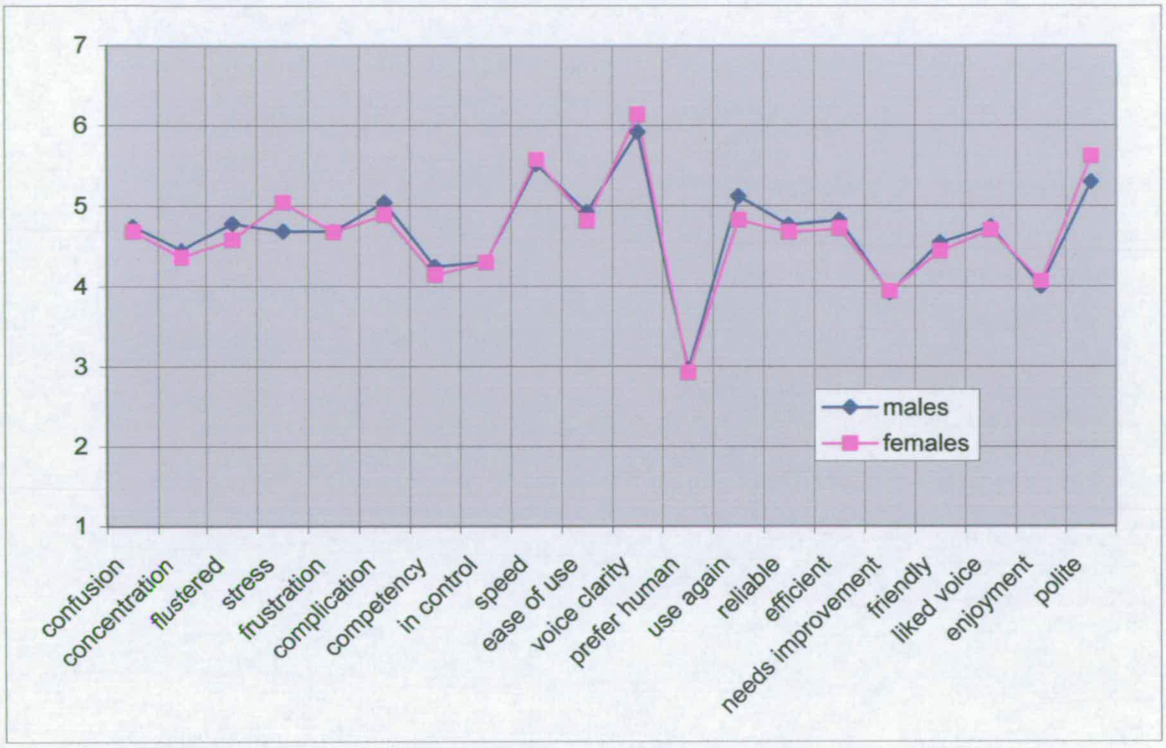


Chart 13. Main scores for UQ0 attributes, split according to gender.

5.5.4 Changes in usability ratings following SID (UQ1-UQ0)

The second set of analyses concerned the impact of the presence of a SID on participants' attitudes towards service usability. This change in attitude was measured using the differential scores, computed by subtracting the UQ0 scores from the UQ1 scores. A univariate ANOVA was then run with age, gender and offer location (three levels: referred to as Welcome, ID&V and SID Location) as between-group variables.

There were no statistically significant differences for main effects of age, gender or SID location. The Intercept value in Table 27 represents the participants' overall change in attitude (positive or negative, compared to 0 which would equal no change). Based on the value obtained [p = .790], it was concluded that the presence of a SID did not have an impact on participants' attitudes towards service usability overall.

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Between-participant effects					
Intercept	.032	1	.032	.072	.790
AGE	.931	2	.465	1.037	.358
GENDER	1.146	1	1.146	2.553	.113
SID_LOCATION	.115	2	.057	.128	.880
AGE * GENDER	.160	2	.080	.179	.837
AGE * SID_LOCATION	.399	4	.100	.222	.925
GENDER * SID_LOCATION	.341	2	.170	.379	.685
AGE * GENDER * SID_LOCATION	.477	4	.119	.266	.899
Error	43.080	96	.449		

Table 27. Univariate ANOVA on differential usability mean scores (UQ1-UQ0).

Differential attitude scores (UQ1-UQ0) were also computed for each of the individual questionnaire statements and Univariate ANOVAs (with the same between-group variables) were performed. Differential score profiles for overall scores (Intercept) and for factors age, gender and SID location are shown in Chart 14, Chart 15, Chart 16 and Chart 17 respectively.

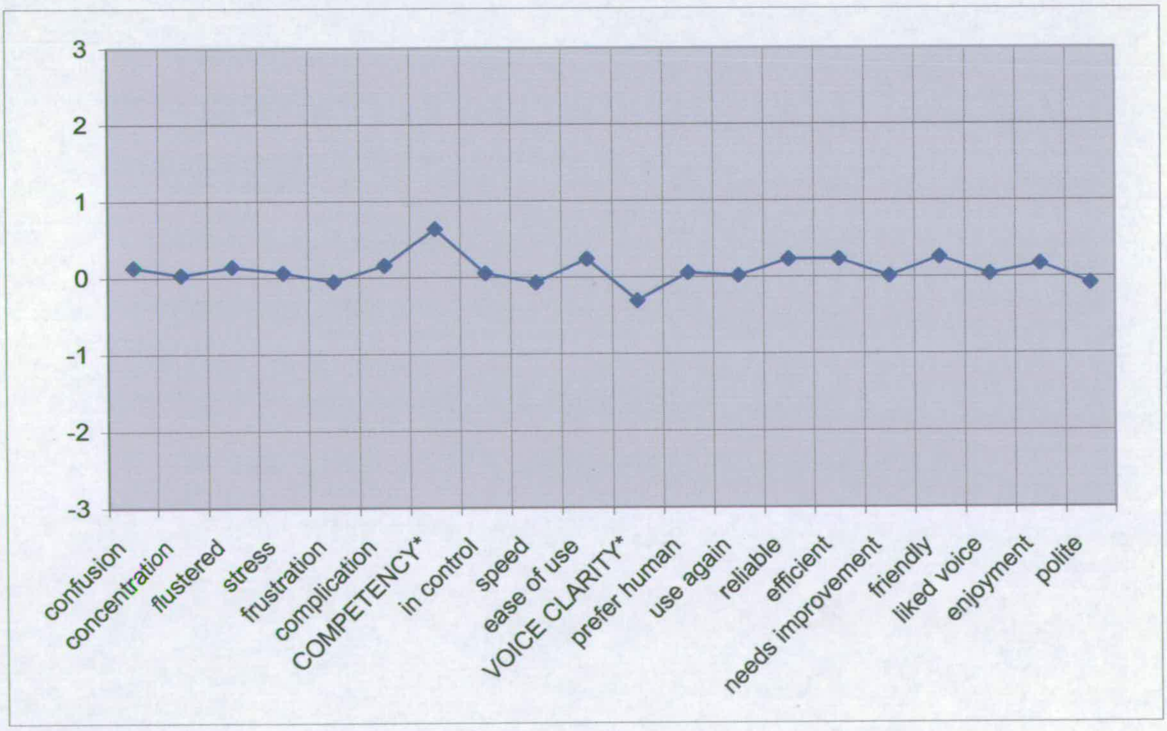


Chart 14. Differential score profiles (UQ1-UQ0), overall scores [*p<.05].

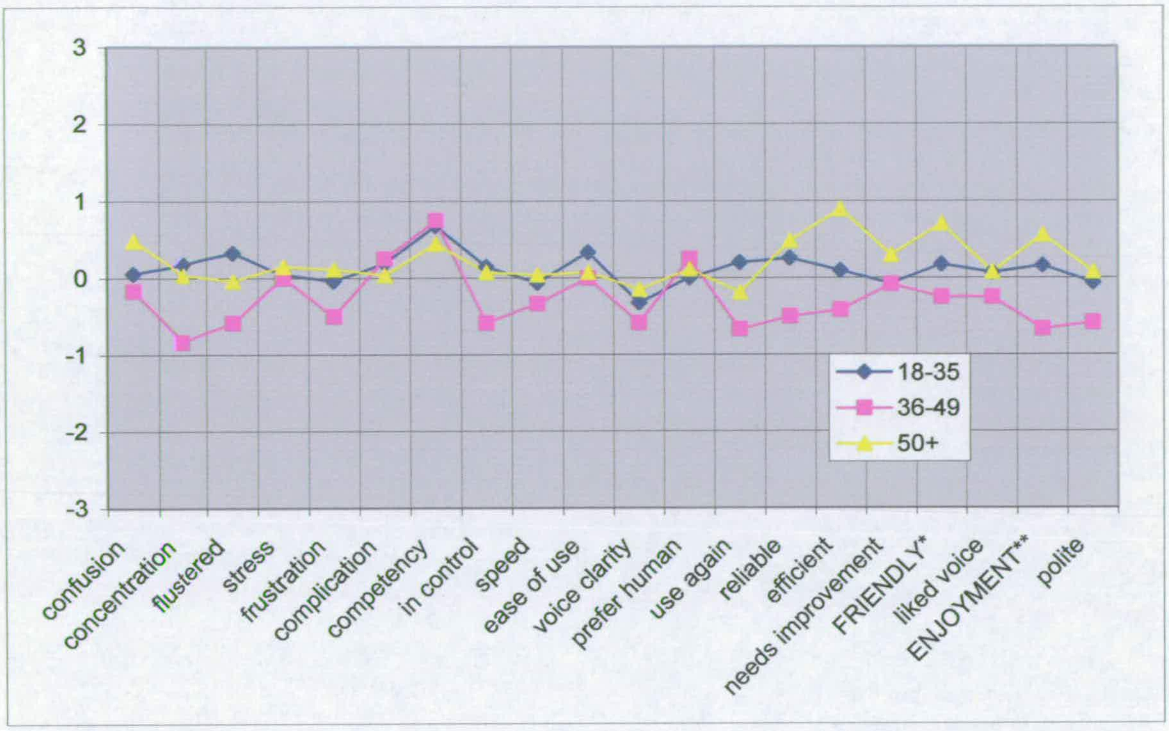


Chart 15. Differential score profiles (UQ1-UQ0), split according to age group [*p<.05; **p<.01].

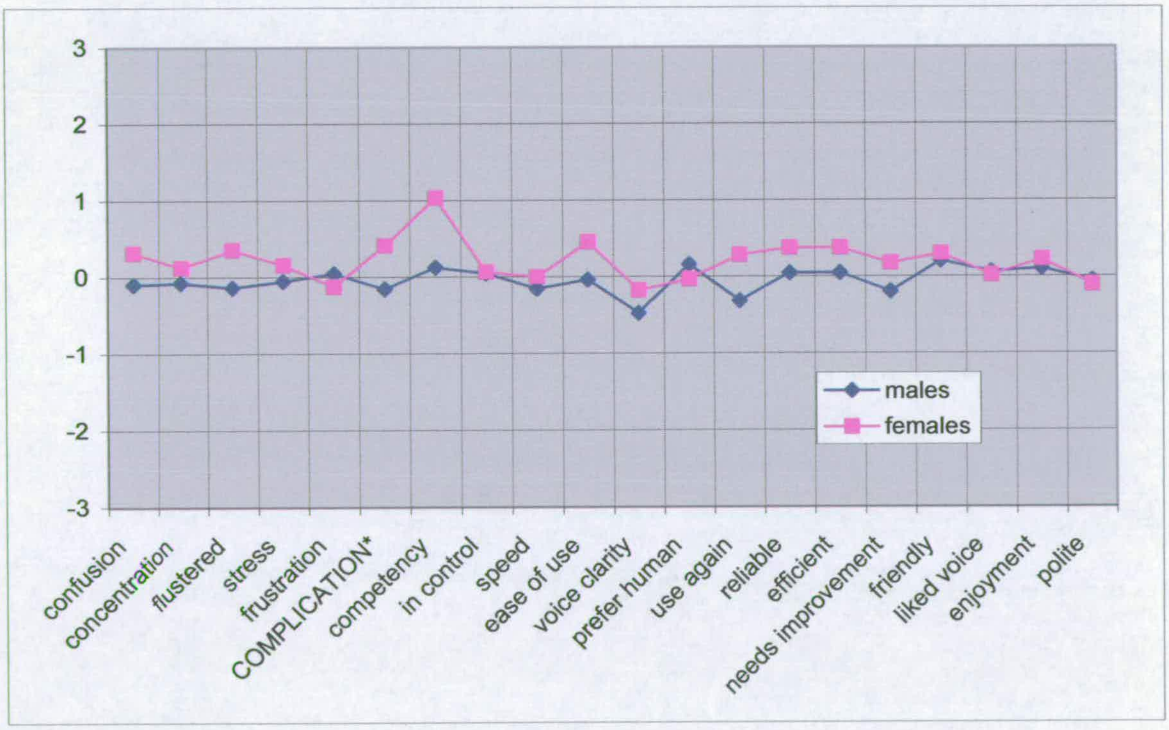


Chart 16. Differential score profiles (UQ1-UQ0), split according to gender [*p<.05].

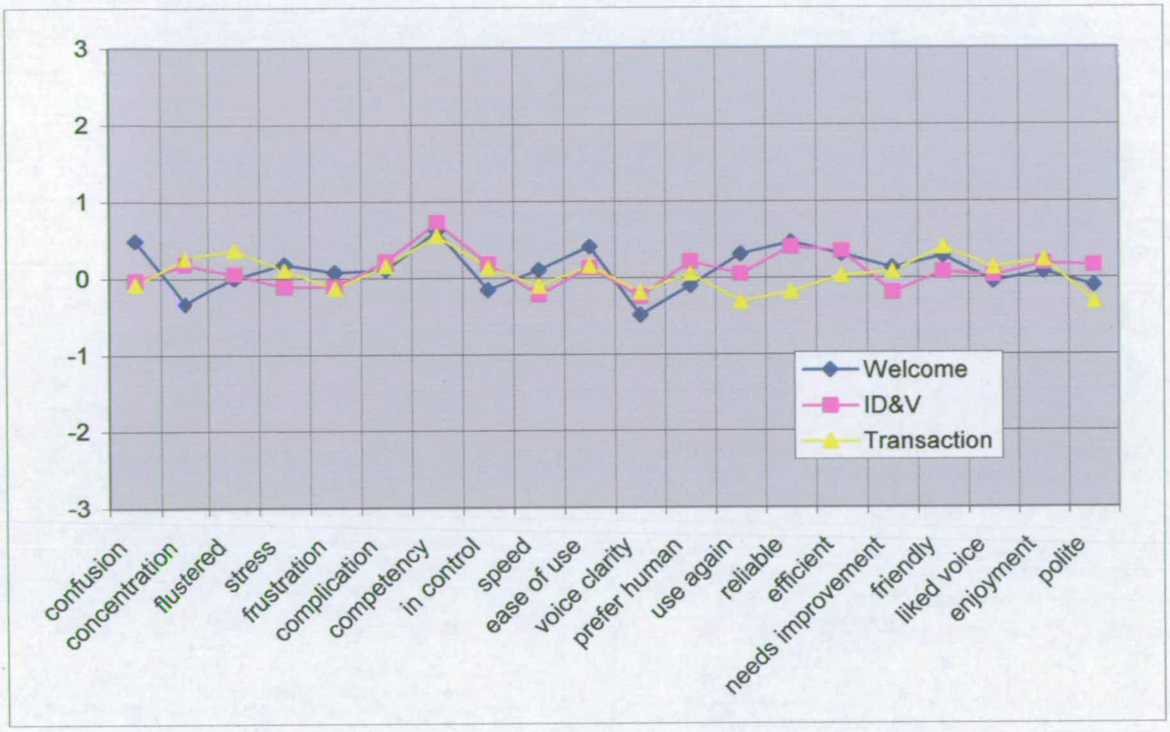


Chart 17. Differential score profiles (UQ1-UQ0), split according to SID location.

There were very few statistically significant differences in the analyses. Overall (Chart 14), participants' attitude towards the usability of the service changed in that they (represented by 'competency' in the profile chart) *knew better what to do* when operating the automated service [$df = 1$, $F = 6.916$, $p = .010$]; however, their attitude towards the *clarity of the voice* had been reduced [$df = 1$, $F = 6.281$, $p = .014$].

Two questionnaire items showed statistical significance in the age group comparison of differential scores (Chart 15): perceived *friendliness* of the service [$df = 2$, $F = 3.660$, $p = .029$] and the degree to which participants *enjoyed using the service* [$df = 2$, $F = 5.174$, $p = .007$]. Post Hoc tests revealed significant differences between the two older age groups in terms of *perceived friendliness* [$p = .043$]; and significant differences for *enjoying using the automated service*, both between the youngest and mid-age groups [$p = .016$] and between the mid-age and oldest age groups [$p = .001$].

Only one questionnaire item turned out significantly different when comparing the change in attitudes between males and females (Chart 16): after the third phone call with the SID offer, female participants found the service *less complicated to use* [df = 1, F = 5.270, p = .024].

There were no significant differences for analyses of differential scores based on SID offer location (Chart 17). Furthermore, there were no significant interactions between factors in the analyses of the differential scores.

5.5.5 Attitudes towards the SID dialogue component (SIDQ)

The current experiment included an additional set of 16 questionnaire statements – referred to here as 'SIDQ' – addressed at capturing participants' attitudes towards the SID offer which occurred during the third call to the service. All participants (N = 114) completed this questionnaire after it had been established that they were all aware that an overdraft offer had been played (there was no control-group in the experiment). A univariate ANOVA was run on the SIDQ mean scores with age, gender and offer location as between-group variables (Table 28).

As for the analysis of UQ0 above, there was a moderate effect for age (questionnaire mean scores: 18-35 M = 4.34; 36-49 M = 4.11; 50+ M = 3.56). Post Hoc tests revealed that the difference between the oldest (with the lowest mean score overall) and the youngest group (with the highest mean score overall) was significantly different [p = .003].

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Between-participant effects					
AGE	8.154	2	4.077	3.843	**.025**
GENDER	.668	1	.668	.630	.429
SID_LOCATION	2.006	2	1.003	.945	.392
AGE * GENDER	.047	2	.024	.022	.978
AGE * SID_LOCATION	1.570	4	.393	.370	.829
GENDER * SID_LOCATION	.200	2	.100	.094	.910
AGE * GENDER * SID_LOCATION	6.325	4	1.581	1.490	.211
Error	101.852	96	1.061		

Table 28. Univariate ANOVA on usability mean scores (SIDQ).

Univariate ANOVAs (with between-group variables age, gender and SID location) were also performed on each of the individual SIDQ statements; score profiles for main factors are shown in Chart 18, Chart 19 and Chart 20.

Six questionnaire items were statistically significant with age as main effect (Chart 18). On the issue of feeling *happy to apply for an overdraft through the service* [*df* = 2, *F* = 4.839, *p* = .010], Post Hoc tests showed that oldest age group took a more negative attitude than both the mid- [*p* = .008] and the youngest [*p* = .024] age groups. The remaining items revealed significant differences mainly between the youngest and oldest participants groups, with the oldest participants taking a more negative attitude to the SID dialogue. The oldest age group significantly *preferred a human agent* to make the overdraft offer [*df* = 2, *F* = 4.772, *p* = .011, Post Hoc = .004], thought it was *harder to understand the overdraft offer* [*df* = 2, *F* = 5.181, *p* = .007, Post Hoc = .001], perceived the information in the offer *less relevant* [*df* = 2, *F* = 3.340, *p* = .040, Post Hoc = .015], found the offer *less appropriate for this kind of service* [*df* = 2, *F* = 4.305, *p* = .016, Post Hoc = .005] and found the method of using offers to give product information as *less efficient* [*df* = 2, *F* = 3.687, *p* = .029, Post Hoc = .011].

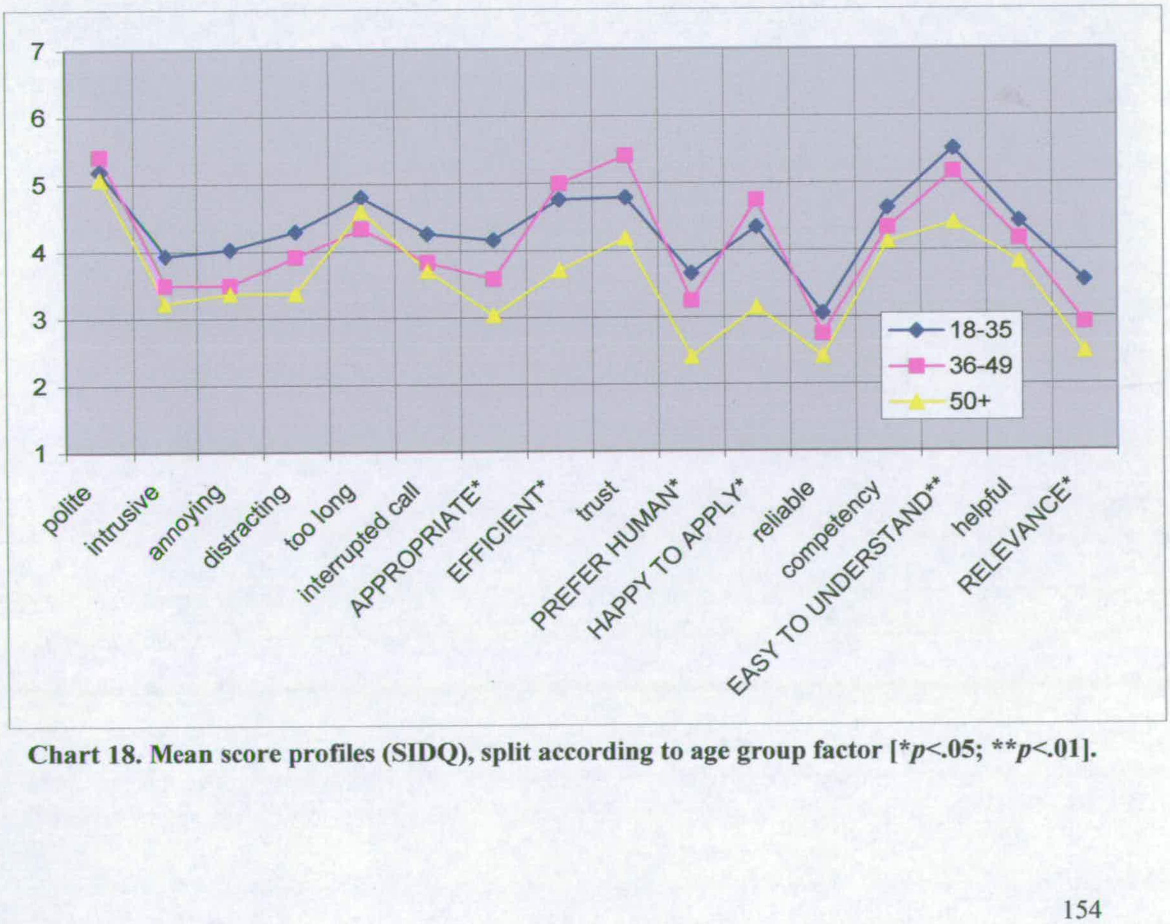


Chart 18. Mean score profiles (SIDQ), split according to age group factor [*p<.05; **p<.01].

There were no statistically significant differences in the analysis based on gender (Chart 19). In terms of proposal location (Chart 20), it can be summarised that the Welcome location generated the most positive responses [$M > 4$, above 'neutral' response] in terms of perceived social characteristics (items *intrusiveness, annoyance, distraction*) and channel suitability (*interrupted the call* and *appropriateness*) compared to the ID&V and Transaction location [$M < 4$, below 'neutral']. However, it was only one questionnaire item, the *appropriateness of the overdraft offer in the service*, which was statistically significant [$df = 2$, $F = 3.900$, $p = .024$]: the Welcome group was significantly more positive toward the offer compared with the ID&V [$p = .002$] and the Transaction [$p = .000$] groups.

There were no significant interactions between factors in the analyses of SIDQ.

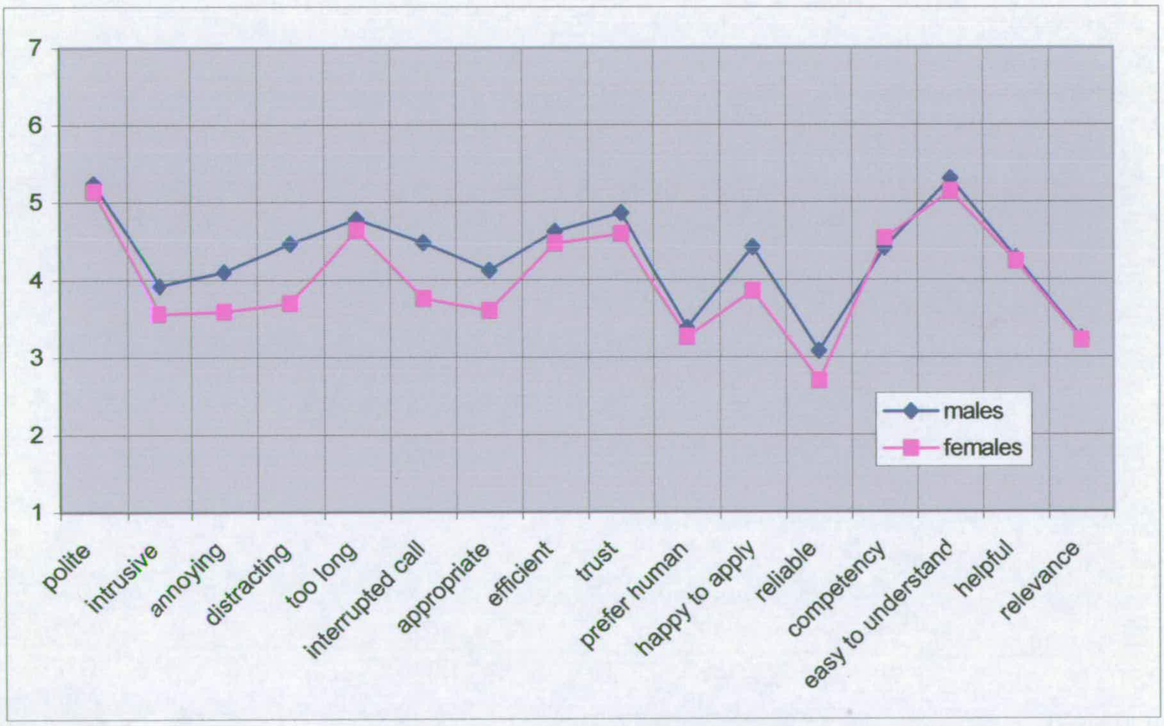


Chart 19. Mean score profiles (SIDQ), split according to gender.

Issues concerning reliability and consistency in the construction of new questionnaires were introduced in Section 3.10.5. Cronbach's Alpha is commonly applied as a method for measuring the average inter-item correlation; if the inter-item correlation is sufficiently high (>.75) it can be concluded that the items measure a single unidimensional latent construct. The Cronbach's Alpha for SIDQ was .90, which suggests a satisfactory level of consistency.

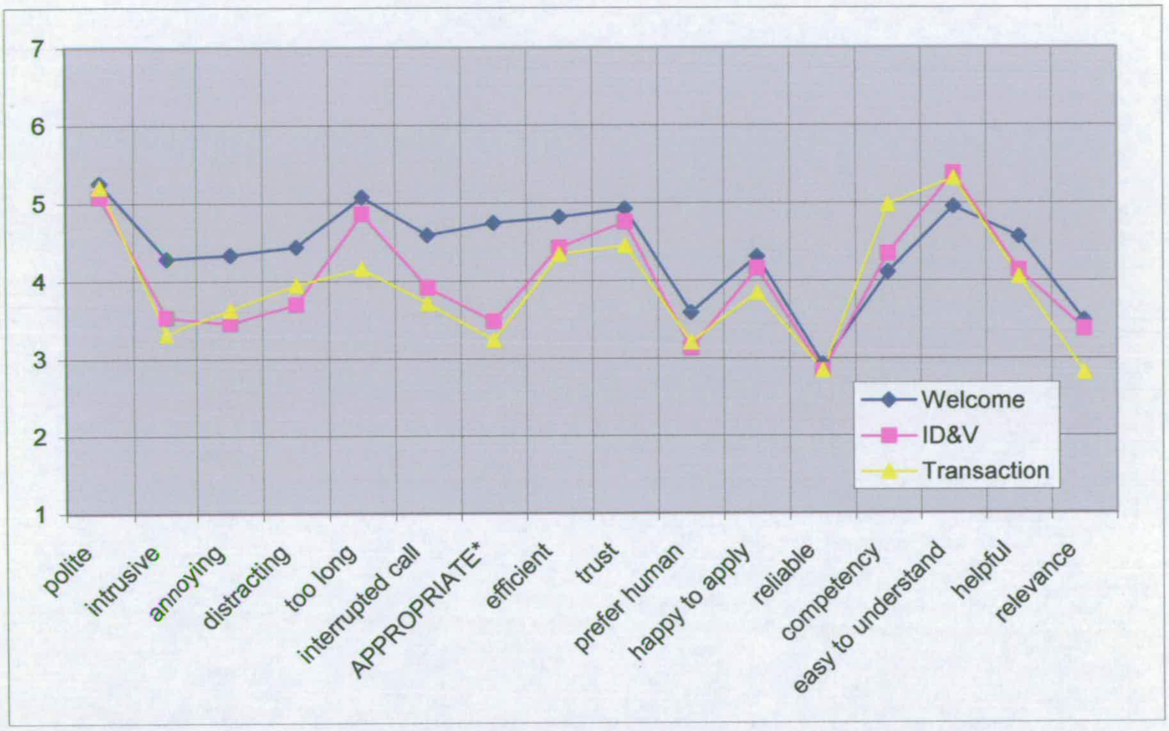


Chart 20. Mean score profiles (SIDQ), split according to proposal location [*p<.05].

5.5.6 Changes in usability ratings after overdraft application task (UQ2-UQ0)

In their fourth phone, participants were instructed to apply for an overdraft. Their attitudes toward the service usability was measured after this phone call (UQ2). To explore the impact of the overdraft application task on participant attitudes to the service, the differential scores were computed by subtracting the baseline usability measure (UQ0) from scores obtained after phone call four (UQ2). Univariate ANOVAs were run on the differential scores (N = 96), excluding the subset of participants in the Welcome group who had been assigned the condition 'not eligible for an overdraft'. Between-participant variables were age, gender and SID location (Table 29). Overall, the drop in attitude to service usability (M = -0.457) was highly statistically significant [Intercept, p = .000].

Univariate ANOVAs were also computed on the differential scores for individual questionnaire items. Profile charts for differential scores based on Intercept (Chart 21), age (Chart 22), gender (Chart 23) and SID location (Chart 24) are shown below.

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Between-participant effects					
Intercept	10.027	1	10.027	14.061	**.000**
AGE	1.315	2	.657	.922	.402
GENDER	1.616	1	1.616	2.267	.136
SID_LOCATION	3.124	2	1.562	2.190	.119
AGE * GENDER	.761	2	.380	.533	.589
AGE * SID_LOCATION	3.169	4	.792	1.111	.357
GENDER * SID_LOCATION	1.437	2	.719	1.008	.370
AGE * GENDER * SID_LOCATION	3.989	4	.997	1.398	.242
Error	55.624	78	.713		

Table 29. Univariate ANOVA on differential mean scores (UQ2-UQ0), for participants who were eligible for an overdraft (*N* = 96).

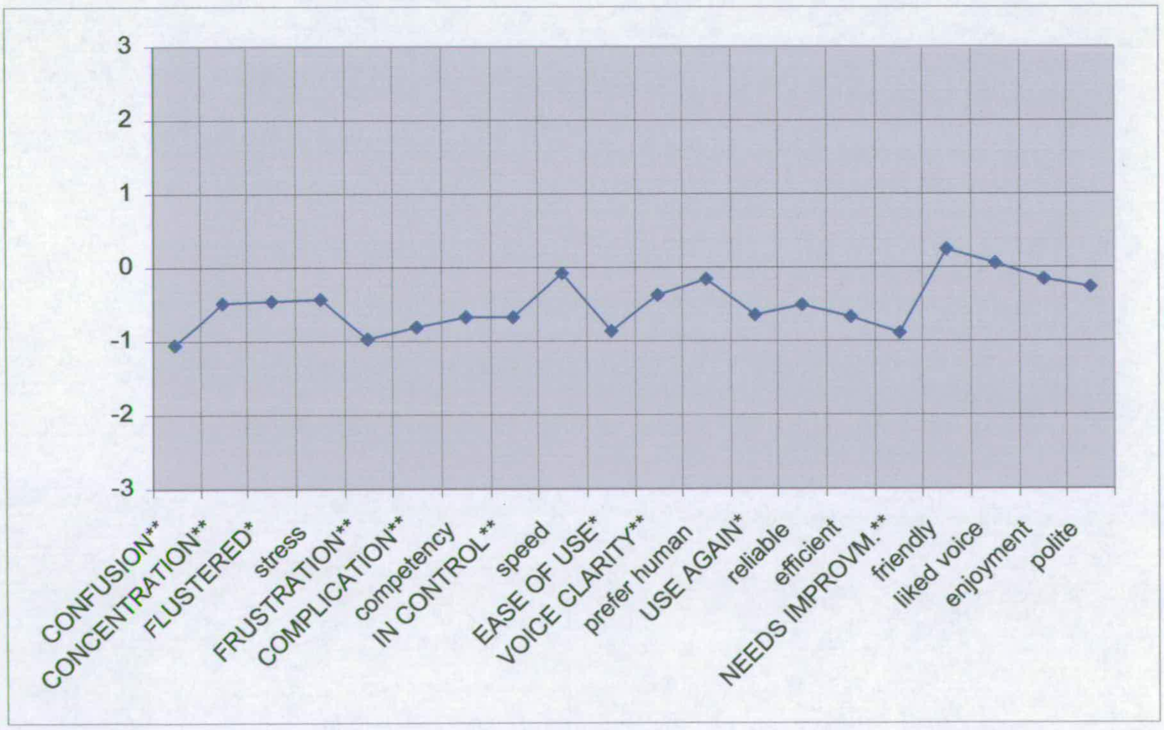


Chart 21. Differential score profile (UQ2-UQ0), for participants who were eligible for an overdraft (*N* = 96), [p*<.05; ***p*<.01].**

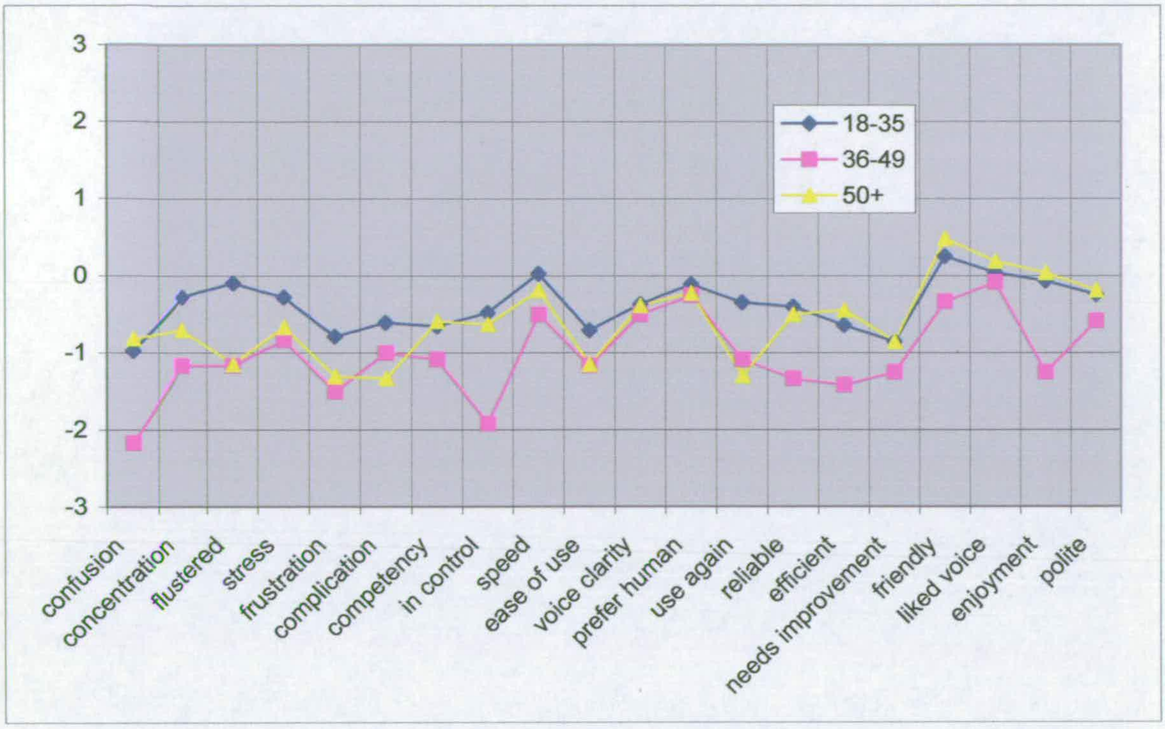


Chart 22. Differential score profile (UQ2-UQ0), split according to age groups for participants who were eligible for an overdraft ($N = 96$). Note, the lack of statistical significance for items such as *in control* are due to small participant sample in the 36-49 age group.

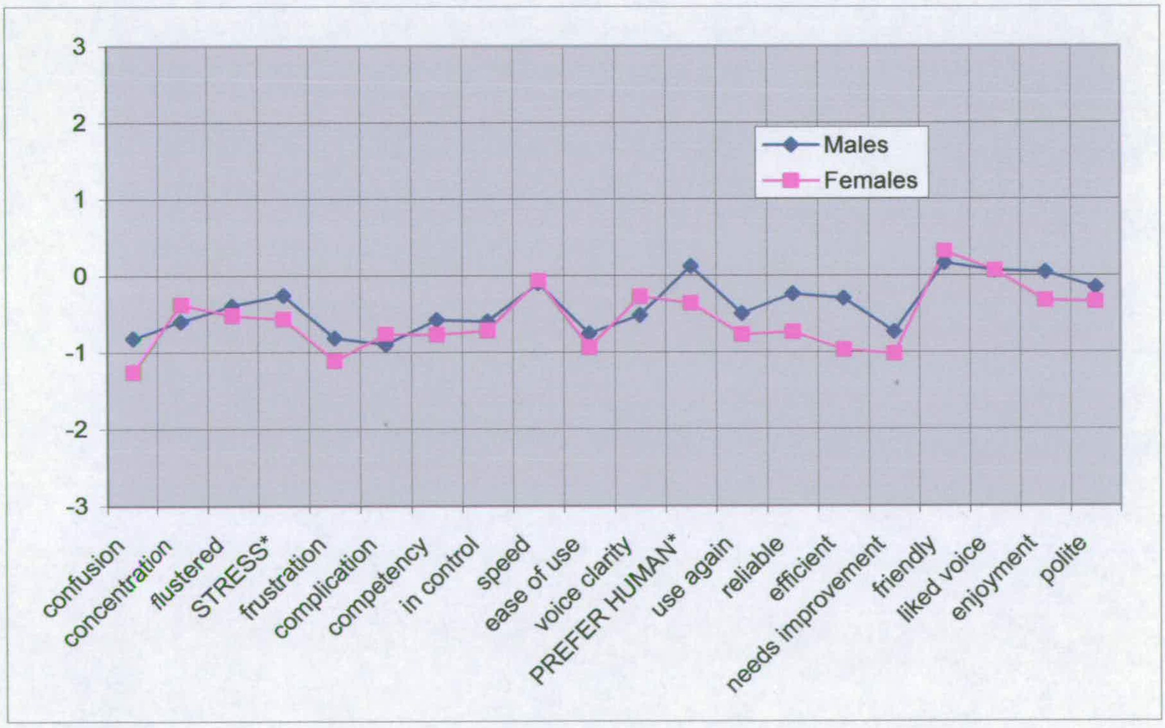


Chart 23. Differential score profile (UQ2-UQ0), split according to gender for participants who were eligible for an overdraft ($N = 96$), [*$p<.05$].

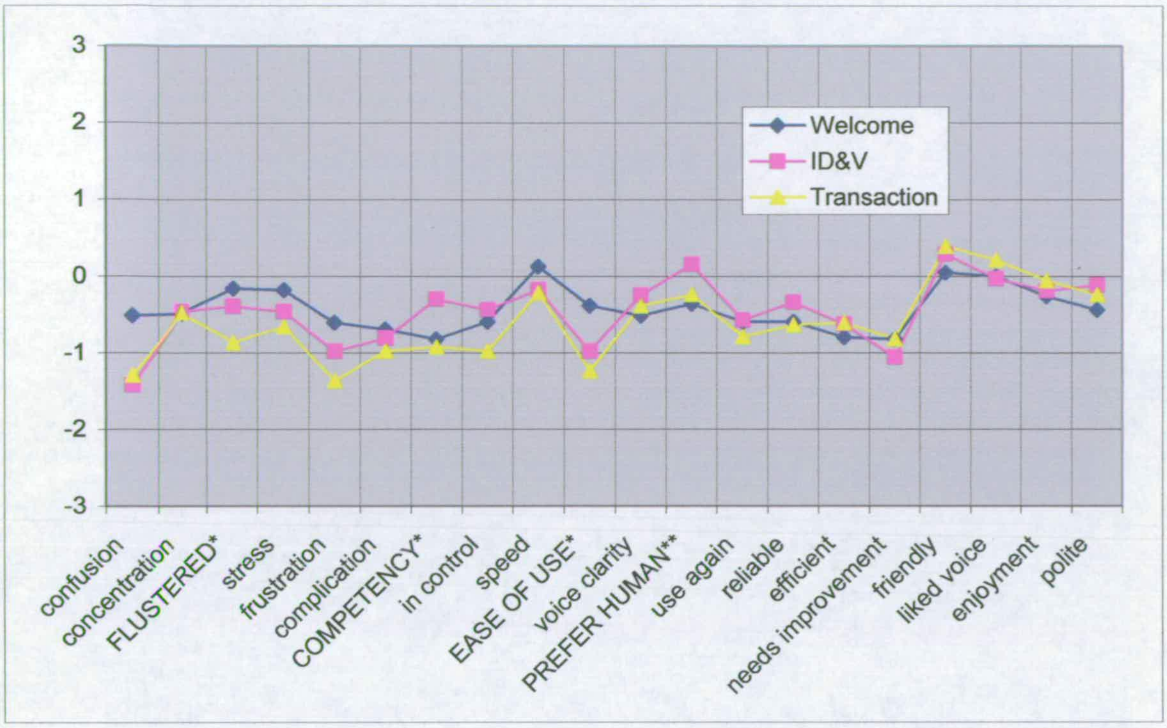


Chart 24. Differential score profile (UQ2-UQ0), split according to SID location for participants who were eligible for an overdraft (*N* = 96), [p*<.05; ***p*<.01].**

The change in attitude was negative and statistically significant for a large number of questionnaire items (Intercept values, Chart 21). In comparison with attitudes after the first two phone calls to the service (UQ0), participants found the service in the fourth phone call more *confusing to use* [*df* = 1, *F* = 16.194, *p* = .000], that they had to *concentrate harder* [*df* = 1, *F* = 7.163, *p* = .009] and felt *more flustered* [*df* = 1, *F* = 4.193, *p* = .044]. Participants also found the service *more frustrating to use* [*df* = 1, *F* = 11.890, *p* = .001], *more complicated* [*df* = 1, *F* = 13.451, *p* = .000] and that they felt *less in control when using the service* [*df* = 1, *F* = 8.485, *p* = .005]. There were significant reductions in the perceived *ease of use* [*df* = 1, *F* = 5.876, *p* = .018] and the *clarity of the voice* [*df* = 1, *F* = 7.474, *p* = .008]. Participants were also less *happy about using the service again* [*df* = 1, *F* = 6.359, *p* = .014] and felt that the service was in *more need of improvements* [*df* = 1, *F* = 7.650, *p* = .007].

There were no statistically significant differences in the analysis of individual items based on age groups (Chart 22). In the gender analysis (Chart 23), two items were statistically significant: female participants felt *more under stress* [*df* = 1, *F* = 6.052, *p* = .016] and would more strongly *prefer to speak to a human* [*df* = 1, *F* = 5.808, *p* = .018] than did male participants.

In the analysis for SID location (Chart 24), there were statistically significant differences in terms of *feeling flustered* [*df* = 2, *F* = 4.727, *p* = .012], perception of *knowing what to do* (competency) [*df* = 2, *F* = 3.231, *p* = .045] and *preference for speaking to a human being* [*df* = 2, *F* = 5.377, *p* = .006]; Post Hoc tests revealed no further statistical differences. Furthermore, there were differences in perceived *ease of use* [*df* = 2, *F* = 4.035, *p* = .021] where the Post Hoc test revealed statistically significant differences between the Welcome and ID&V groups [*p* = .035] and between the Welcome and Transaction group [*p* = .007].

A number of interaction between factors were significant; these are summarised in Table 30. In addition, there was also a three-way interaction for the perceived *politeness* of the automated service (AGE*GENDER*SID_LOCATION, *p* = .033).

Questionnaire item	Interaction	*p*-value
I felt flustered when using the service.	AGE*SID_LOCATION	.043
I felt under stress when using the service.	AGE*GENDER	.006
When I was using the service I always knew what I was expected to do.	AGE*SID_LOCATION	.047
I would prefer to talk to a human.	AGE*SID_LOCATION	.044
	GENDER*SID_LOCATION	.006

Table 30. Summary of significant interactions for differential scores (UQ1-UQ0), for participants who had experienced a SID offer and were eligible for an overdraft (*N* = 96).

Task completion rates in Section 5.5.2 revealed that a significant subset of participants (36.8%) failed to request an overdraft. Differential scores for participants, split according to whether they had said "overdraft" or not, are shown in profile Chart 25; participants who failed to complete the overdraft request took a more negative attitude to the usability of the service.

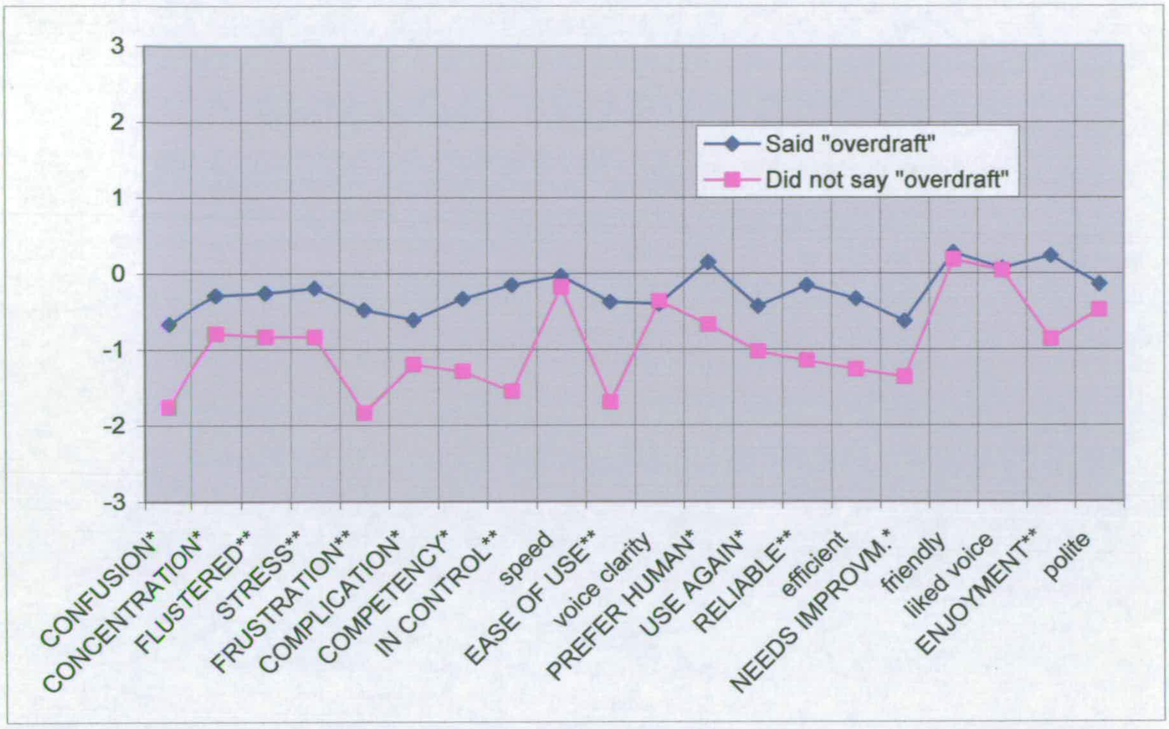


Chart 25. Differential score profile (UQ2-UQ0), for participants eligible for an overdraft (*N* = 96), split according to whether they had said "overdraft" or not [$*p<.05$; $**p<.01$].

In brief, ANOVAs (with age, gender and SID location) run on overall scores showed that the change in attitude (Intercept) within the 'said overdraft' group was not significant [$N = 56$, $df = 1$, $F = 2.338$, $p = .134$]; the change in attitude within the 'did not say overdraft' group was highly significant [$N = 37$, $df = 1$, $F = 61.638$, $p = .000$].

Overdraft task completion had a direct impact on the dependent variable – but was not part of the main experiment manipulation. The relationship between the independent variables in this causal model is illustrated in Figure 18. The ANOVA (with specified factors age, gender and SID location) analyses the *total* effect of the independent variables on attitude scores; this relationship between variables is illustrated (all arrows) in Figure 18.

The ANOVA can be extended to include overdraft task completion as a covariate (ANCOVA); this analysis controls for the effect of the covariate by adjusting the dependent variable scores to what they would be if everyone had the same result on the covariate; this is illustrated in Figure 18 where the effect of the covariate has been removed (dashed arrows). An ANOVA is then applied on the adjusted scores and the *direct* effect of the independent variables (age, gender and SID location) on attitude scores (solid arrows) is obtained. ANCOVA increases the power of an *F*-test by removing unsystematic variance (noise) in the dependent variable; using covariates can show larger effects or can result in effects being eliminated.

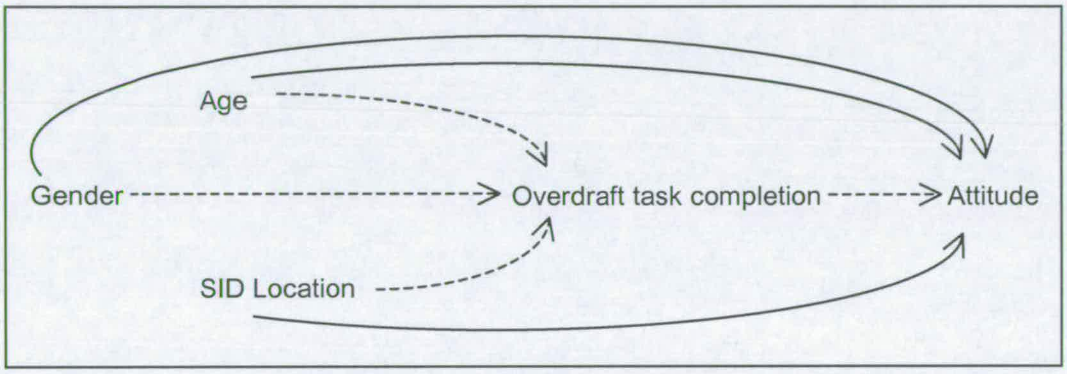


Figure 18. The effect of the independent variables (age, gender and SID location) on attitude, analysed into effects mediated by overdraft task completion (dashed arrows) and direct effects when task completion is controlled for (solid arrows).

To conduct an ANCOVA, the 'assumption of homogeneity of regression slopes' needs to be tenable; that is, the relationship between the dependent variable and the covariate needs to be the same for all groups of participants. To achieve this, an ANCOVA was run with a customised model which tested the interaction between the covariate and the independent variables (full factorial customised model). Results from this analysis showed that there were no significant interactions between the covariate (overdraft completion) and the independent variables; thus, the assumption of homogeneity of regression was satisfied and the ANCOVA could be performed on the data with overdraft completion specified as the covariate.

The overall results from the ANCOVA (independent variables: age, gender, proposal and overdraft completion as covariate) are presented in Table 31. The significance value obtained for the covariate (OD_COMPLETION) clearly indicates that the covariate significantly predicts the dependent variable.

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Between-participant effects					
Intercept	22.593	1	22.593	40.402	**.000**
OD_COMPLETION	12.566	1	12.566	22.471	**.000**
AGE	1.418	2	.709	1.268	.287
GENDER	.094	1	.094	.169	.682
SID_LOCATION	2.858	2	1.429	2.556	.084
AGE * GENDER	.999	2	.500	.894	.413
AGE * SID_LOCATION	2.073	4	.518	.927	.453
GENDER * SID_LOCATION	.082	2	.041	.074	.929
AGE * GENDER * SID_LOCATION	2.600	4	.650	1.162	.334
Error	43.059	77	.559		

Table 31. Univariate ANCOVA on differential mean scores (UQ2-UQ0), for participants who were eligible for an overdraft ($N = 96$).

ANCOVA analyses were also performed on individual differential scores; the significance values with the covariate included are marked in Chart 25[30]. The inclusion of the covariate (compared with the original ANOVA displayed in Chart 25) resulted in a further five items showing significant differences (Intercept) which are summarised here: *perceived stress* [p = .001], *knowing what to do* (competency) [p = .002], *reliability* [p = .003], *efficiency* [p = .013] and *enjoying using the service* [p = .001]. Remaining items which were significant in the original ANOVA analysis remained significant in the ANCOVA analysis – their p-values now all at the higher level of significance [$p < .01$] (except *clarity of the voice* where the p-value was at the lower significance, $p < .05$, but was non-significant in the original analysis).

In the analysis based on the gender factor, the two items (*perceived stress* and *preference for speaking to a human*), which were significant in the ANOVA analysis (Chart 23), were reduced to non-significance in the ANCOVA analysis. This could

[30] The findings regarding the impact of overdraft completion are secondary to the main focus of the research and are therefore presented in abbreviated form, omitting F-values. It will suffice to conclude here that overdraft completion rate had a significant impact on user attitudes.

be explained by the fact that significantly more males than females succeeded with the overdraft request and the inclusion of the covariate cancelled out this difference.

Furthermore, the questionnaire item *knowing what to do* (competency), significant for the SID location factor in the ANOVA (Chart 24), was non-significant in the ANCOVA analysis. Items *felt flustered* and *service easy to use*, at the lower level of significance in the ANOVA analysis (p<.05), were highly significant (p<.01) in the ANCOVA analysis whereas the reverse change in significance levels were obtained for item *prefer to speak to a human*.

In the interactions for factors specified in Table 30, *perceived stress* and *feeling flustered* remained at the same significance level in the ANCOVA. The interactions for *knowing what to do* and *prefer to speak to a human* (AGE*SID_LOCATION interaction) became non-significant; the interaction (GENDER*SID_LOCATION) for *prefer to speak to a human* remained highly significant.

5.5.7 Impact of eligibility for overdraft (UQ2-UQ0)

The issue with having to turn down potential applicants was modelled by including an 'ineligible for overdraft' group in the Welcome SID location condition. Subsequently, when these participants requested an overdraft they were refused. The impact of having to turn down an applicant was evaluated by running univariate ANOVAs on the differential scores from participants who experienced the Welcome SID proposal and who had also said "overdraft" (N = 28). Between-participants variables were: age, gender and overdraft eligibility (two levels, N = 15 in the 'overdraft allowed' group and N = 13 in the 'overdraft refused' group). With such a low number of participants, some treatments had counts of '0'; the GLM sums of squares (described in Section 3.10.4) were adjusted accordingly to Type IV in order to account for missing cells.

Source	Type IV Sum of Squares	df	Mean Square	F	Sig.
Between-participant effects					
Intercept	3.857	1	3.857	5.680	**.028**
AGE	2.547(b)	2	1.274	1.876	.181
GENDER	.055(b)	1	.055	.081	.779
OD_ELIGIBLE	2.819(b)	1	2.819	4.153	.056
AGE * GENDER	1.961(b)	1	1.961	2.889	.106
AGE * OD_ELIGIBLE	1.061(b)	2	.530	.781	.472
GENDER * OD_ELIGIBLE	.008(b)	1	.008	.011	.916
AGE * GENDER * OD_ELIGIBLE	.000	0	.	.	.
Error	12.900	19	.679		

Table 32. Univariate ANOVA on differential mean scores (UQ2-UQ0), for participants in the Welcome group who had said "overdraft" (*N* = 28).

For the participant group as a whole (Table 32), the overall change in attitude (UQ2-UQ0) was moderately significant [p = .028]. The impact of overdraft eligibility (OD_ELIGIBLE) approached significance [p = .056]: the overdraft eligible participant group rated the service usability more positively overall (M = 0.1340) compared to the overdraft ineligible group (M = -0.7253).

On a per questionnaire statement basis there was statistical significant drop in attitude (Intercept) for: *feeling the service was complicated* [df = 1, F = 8.824, p = .008], *knowing what to do* [df = 1, F = 5.271, p = .033], *feeling in control* [df = 1, F = 5.767, p = .027], *clarity of the voice* [df = 1, F = 4.712, p = .043], *service reliability* [df = 1, F = 5.663, p = .028], *feeling the service needs improvements* [df = 1, F = 10.780, p = .004], *efficiency of service* [df = 1, F = 4.469, p = .048] and *politeness of the service* [df = 1, F = 5.193, p = .034].

For main effects, there was significant difference in perceived *reliability of the service* [df = 1, F = 10.780, p = .004] where eligible applicants were more positive towards the service (M = 0.4667) compared to ineligible applicants (M = -1.1538). Furthermore, ineligible applicants (M = -1.4615) found that the service *needed improvement* more than eligible applicants did (M = 0.1333), [df = 1, F = 8.382, p = .009].

5.5.8 Overdraft application task: utterance data

At each stage in the dialogue, system log files were used to record information about participants' interaction with the service. The log files contained details of the system prompt played, the error level, the recognition result returned by the system and whether the participant responded with voice or touch-tone button input. This data was then used to obtain information about each participant's navigational route through the system.

This section is concerned with participants' navigational route through the service up to the point of succeeding (or failing) to complete the overdraft request task. For this purpose, only the log data after the participant had completed the balance and statement order tasks in the fourth phone call will be considered. Thus, the analyses described in this section start at the dialogue stage where the caller is faced with the system prompt "would you like another service?" (Figure 4) and the task at hand at this point is to apply for an overdraft on the current account.

At this point in the dialogue (Table 33), seven out of the 114 participants requested an overdraft: four of these participants said "overdraft", while three participants also specified the name of the account (e.g. "overdraft on my current account"). All seven participants had volunteered "statement" in at least one of their four phone calls to the service. One further participant (from the 'not volunteered' category) made a mistake at this point in the dialogue and answered "no" which consequently ended the phone call; the remaining 106 participants responded "yes".

	Volunteered "statement" at least once (*N* = 39)		Never volunteered "statement" (*N* = 75)	
Recognition result	**completed overdraft**	**failed overdraft**	**completed overdraft**	**failed overdraft**
"overdraft"	7	-	0	-
"yes"	26	6	39	35
"no"	0	0	0	1
TOTAL	**33**	**6**	**39**	**36**

Table 33. Participant responses to "Would you like another service", (*N* = 114).

The "yes" response subsequently triggered the system prompt "Please select balance, recent transactions or another service" (Main Menu 'a' in Figure 4). At this point in the dialogue (Table 34), 26 participants said "overdraft": the majority of these (*N* = 19) had volunteered the "statement" keyword in at least one of their phone calls (two participants in the volunteer group had also stated the name of the account). The remaining 80 participants responded with "another service" which triggered Main Menu 'b': "In addition you can select funds transfer, item search, order statement or change TIN. Which service would you like?". Participant responses to this prompt are detailed in Table 35 below.

	Volunteered "statement" at least once (*N* = 32)		Never volunteered "statement" (*N* = 74)	
Recognition result	**completed overdraft**	**failed overdraft**	**completed overdraft**	**failed overdraft**
"overdraft"	19	-	7	-
"another service"	7	6	32	35
TOTAL	**26**	**6**	**39**	**35**

Table 34. Participant responses at Main Menu 'a' stage, (*N* = 106).

In total, 21 participants requested "overdraft" at this point, while a further six participants requested other items from the main menu options (five said "item search" and one "balance"). A further 18 participants said something which was categorised as invalid utterances; of these, 13 utterances were rejected by the system (85% of these utterances were of the category "none" or "none of these"), three utterances were misrecognised, and there were two cases of "another service". The remaining 35 participants remained silent at this point; silences constituted a significant part (46.3%) of the responses given by participants who had not volunteered "statement" in any of their phone calls to the service.

	Volunteered "statement" at least once (*N* = 13)		Never volunteered "statement" (*N* = 67)	
Recognition result	**completed overdraft**	**failed overdraft**	**completed overdraft**	**failed overdraft**
"overdraft"	3	-	18	-
Other menu item	0	1	3	2
Invalid utterance	3	2	4	9
Silence	1	3	7	24
TOTAL	**7**	**6**	**32**	**35**

Table 35. Participant responses (*N* = 80), after hearing all service options in the menu (Main Menu 'a' and 'b').

5.5.9 Overdraft application task: navigational route length

This section details the total navigational route length for participants up to the point of either succeeding in completing the overdraft request (Figure 19), or failing to do so (Figure 20). The route length data complement the utterance analysis from the previous section by exploring the total number of turn-taking iterations involved in completing or failing/giving up on the overdraft task. For the purpose of this analysis, each system-prompt-user-response sequence was treated as one unit and received a score of '1' for navigational route length. The main reason for looking at the total route length (as opposed to actual path through the dialogue) is to reduce the number of route permutations for participants who persisted with looping back to the menu or trying alternative service options in order to find the overdraft. To recapitulate, only responses following the completion of the balance and statement order tasks will be included in the analyses, at the point where the caller is prompted with "would you like another service?" and the task at hand is to apply for an overdraft. Responses to this system prompt are represented by the route length of '1' in Figure 19 and Figure 20.

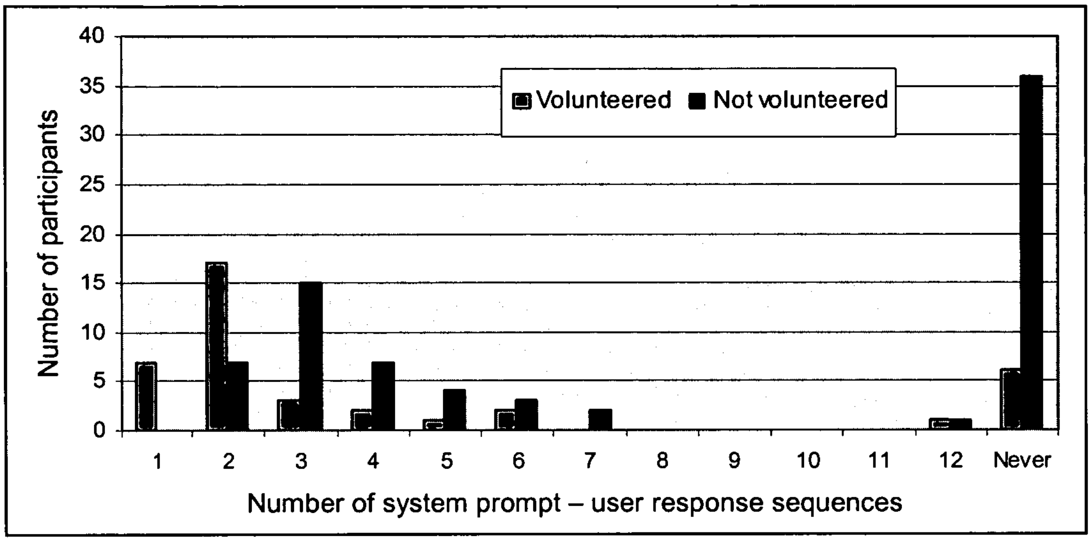


Figure 19. The length of the navigational route for participants who succeeded in completing the overdraft task, split according to their volunteering behaviour in the statement task. For comparison, participants who did not complete the overdraft task are included in the 'Never' category, far right.

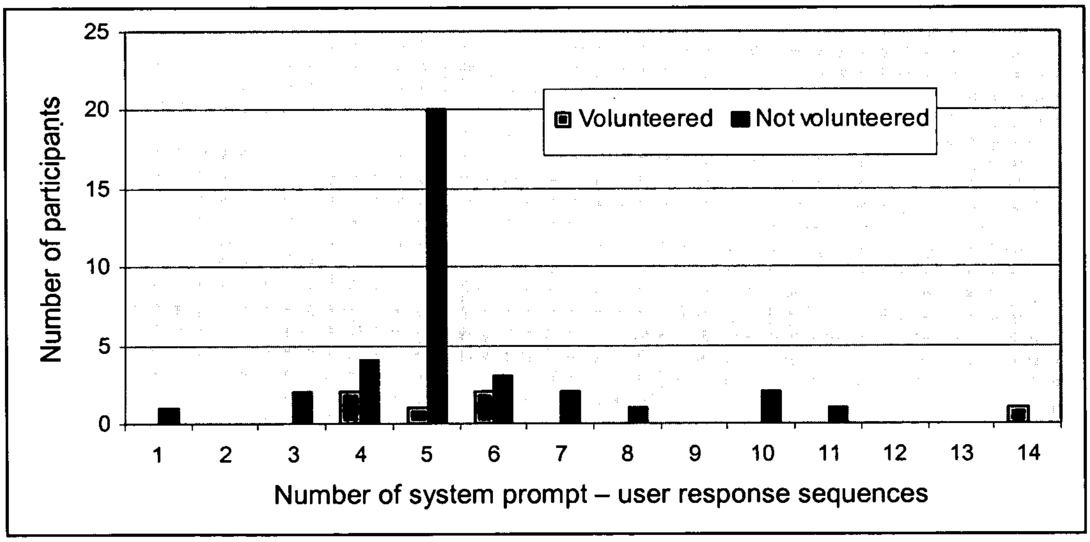


Figure 20. The length of the navigational route for participants who failed to complete the overdraft task, split according to their volunteering behaviour in the statement task. Their phone call either ended by hanging up, or by getting transferred to a human operator (call breakout).

Figure 19 represents the route length by participants who completed the overdraft task, split according to whether or not they had volunteered "statement" in any of their calls. A route length of '1' means that participants responded "overdraft" to the system prompt "would you like another service?". A route length score of '2' means the participant had said "yes" to "would you like another service?", and then said "overdraft" at the Main Menu 'a' stage. Participants with a route length of '3' had said "another service" at Main Menu 'a' and then "overdraft" at Main Menu 'b'. Most of the participants in group '4' had, before saying "overdraft", a silence or reject and therefore had to go through an extra prompt-response sequence in the error recovery. The majority of participants with scores of '5' or over tried at least one other service before saying "overdraft". For comparison, navigational routes for participants who failed to complete the overdraft task are also included ('Never' category in Figure 19). There was a tendency among participants in the volunteered statement group to request the overdraft before they had heard all the options in the main menu.

The same criterion for calculating the route length described in the previous section was used in Figure 20, although instead of saying "overdraft" these participants had put the phone down or their call was transferred to a human agent. All but one participant (who had made the mistake of answering "no" in response to "would you like another service?") chose at least to hear both Main Menu 'a' and 'b' before their call ended. At the point of having listened to all the menu options, the majority of participants stayed silent or tried out the "item search" service option listed in Main Menu 'b'. In the real service, the item search option enables customer to search for a transaction on their account, either by giving the amount or cheque number. Participants, however, had not received priming about the functionality of any of the service options and so their assumption that 'item search' might help them with their task was a reasonable conclusion.

5.5.10 De-briefing interview feedback

All participants took part in a structured interview after their experience with the automated service. All but one participant were aware that an overdraft proposal had been played in the third phone call to the service. Some 63.2% (N = 72) of participants claimed that they could see some benefits with having proposals in this kind of automated service, 29.8% (N = 34) did not perceive any benefits and the remaining participants responded that they did not know.

When asked if, using the automated service as a real service, they would prefer never to be offered an overdraft 45.6% (N = 52) of participants answered "yes". Excluding the "don't know" participant responses, the preference not to hear the SID offer in the individual location groups was: 32.4% (N = 12) for Welcome, 54.1% (N = 20) for ID&V and 54.1% (N = 20) for Transaction (non-significant). Participants (N = 52) who said that they would prefer never to be offered an overdraft were also asked whether the presence of a proposal would discourage them from using the service in the future; 36.4% (N = 18) responded that it would.

The exit interview continued by asking participants (N = 42) who had not said "overdraft" during their fourth phone call the following question: *"Why did you not apply for the overdraft?"*. 40 participants responded to this question. One participant stated he had made a mistake and had answered "no" to the system prompt "would you like another service?". Out of the remaining 39 participants, the majority (79.5%) stated that they did not say "overdraft" as there was no such option in the main menu. Examples of actual participant responses from this group were: "I didn't get the option in the main menu" and "I expected the list of options to include an overdraft option".

The comments from the remaining eight participants (20.5%) reflected interpretation problems beyond the issue of the main menu listing. Four of these participants stated that they had not completed the overdraft application because they did not understand or remember how to: "Couldn't remember how to, was expecting to select it from the menu"; "Didn't understand how to apply for the overdraft. It was not obvious from the menu how to go about doing it"; "Wasn't anything on the menu regarding overdraft. Didn't know what the keyword was to get information on overdraft, so I stayed silent"; and "Couldn't remember how to. Don't think I heard overdraft message in the menu". Their comments indicate that they were aware that the option should be selectable at the menu but that they lacked the precise instructions for how to accomplish this. Another participant stated that: "I didn't get the overdraft message after the balance and couldn't remember what the third call message said about how to get an overdraft. Tried another service option but it didn't help". The remaining three participants expressed more general confusion: "It was confusing"; "I had forgotten what to say"; and "Didn't understand it. Got transferred to an operator".

5.6 Conclusions

The research in this chapter centred around four themes:

1) presence/absence of SID

2) attitudes to the contrasting SID locations

3) the impact of having to turn down ineligible applicants in the Welcome SID location group and

4) participants' ability to locate and select the hidden overdraft option.

A further three issues were identified during the experiment research:

5) the effect of age on attitudes to the overdraft offer,

6) the impact of overdraft task completion on participant attitudes and

7) participants mental models of menu-driven automated telephone services.

5.6.1 The impact of presence of SID

The prediction for this experiment was that the presence of a SID delivery in the dialogue would have little or no impact on user attitudes toward the usability of the service. All participants experienced a SID offer and the analysis of the differential scores (UQ1-UQ0) overall revealed no significant difference (Intercept) [*p* = .790]. Only two individual statements (*knowing what to do* and *voice clarity*) showed statistically significant differences. These results reconfirm the findings from the previous research (Experiment 1): the SID offer did not have a negative impact on users' attitudes toward the automated service.

As with Experiment 1, around 40% of participants responded that – in real life use of the service – they would prefer never to be offered an overdraft; however, the majority of participants (63.2%) did perceive some benefits with receiving such offers.

In sum, in accordance with experiment prediction 1, the delivery of an overdraft offer had no significantly negative impact on user attitudes to the automated service.

5.6.2 The impact of SID location

It was predicted that varying the location of the SID offer would have little impact on participants' attitudes toward the service. This prediction was affirmed by looking at the differential scores (UQ1-UQ0) with the effect of SID location: there were no significant differences overall [*p* = .880] or for individual questionnaire items.

The current experiment included a set of 16 questionnaire statements (SIDQ) aimed at capturing participants' attitudes towards the SID dialogue directly (rather than their attitude towards the usability of the service). It was predicted that the Transaction SID location could be perceived as more intrusive or distracting, due to its nested position within the dialogue. Although the Welcome SID offer scored more positively (compared with ID&V and Transaction offer location) for items such as *intrusiveness*, *annoying* and *interrupted the call* there were only one item that was statistically significant: *perceived appropriateness of the proposal.* The conclusions based on these findings are therefore that varying the SID location has little or no impact on user attitudes toward the offer itself.

Welcome SID location participants were somewhat more positive – when considering real use of the service – towards the idea of being offered an overdraft compared to the ID&V and Transaction location groups (although the differences were not statistically significant). However, participants in the Welcome location group failed to realise that the same offer would be heard at the start of each phone call and that this could become annoying. Participants' failure to recognise this adverse design feature may simply be put down to the fact that they were not asked to reflect on whether or not this would be a problem with repeat use.

To conclude, the current findings do not provide any supporting arguments for selecting one SID delivery location over another. The dialogue engineer has to consider the relative prominence of the information in the offer; it may be suitable to locate an important SID at the ID&V stage of the dialogue, whereas less important information could be postponed until the caller has finished the primary tasks. If located at the end of the call, the caller may hang up before hearing the SID – this location may therefore be unsuitable for more critical messages, but could offer improved perceived usability as the caller can decide whether or not to listen to the message.

In sum, in accordance with experiment prediction 2, contrasting locations for the overdraft offer did not have a significant positive impact on user attitudes to the service. Furthermore, there was no evidence that participants fount the Transaction location for the offer more intrusive than the Welcome or ID&V locations. There was some indication (exit interview) that participants in the Welcome location group were more favourably disposed toward receiving offers in real use of the service however, when considering repeat use of the automated service and having to potentially turn ineligible applicants down, this location was found to be less suited to these kinds of offers.

5.6.3 The effect of age on attitudes to the SID offer

The analysis of SIDQ responses (attitudes toward the SID dialogue) showed that participants in the different age groups responded differently overall [p = .025]: the oldest age group took a more negative attitude to the SID. This effect also carried over to a number of individual questionnaire items: *appropriateness* and *efficiency* of the offer method; the *preference of being approached by a human* with the offer; and the willingness to *apply for an overdraft through the service*. The oldest age group also found the offer information more *difficult to understand*; they also found more strongly that the proposal *was irrelevant* which could account for some of the negative attitude permeating their responses overall.

These findings are interesting and suggest that different age groups may react differently to receiving SID offers; that this is something that should be taken into account when deploying product offers in the automated service. In conclusion, the oldest age group appeared reluctant to both receive an offer and apply for an overdraft through the automated service; in these instances it may be more appropriate to offer a transfer to a human agent for further help and information. Furthermore, it needs to be established whether or not the relevance or comprehension of a particular product (such as an 'overdraft') applies across all customer age categories.

Results from the baseline evaluation of service usability (UQ0) showed that the oldest participant group took a more negative attitude towards the service overall (statistically significant compared to the youngest age group). This means that some of the difference in attitudes overall between the age groups could be attributed to a carry-over effect due to the oldest participants on the whole taking a more negative attitude towards the 'idea' of automated telephone services. In this particular experiment, the mid-age group in particularly was underrepresented and further experiments are necessary to explore the significance of differences in age groups further.

5.6.4 Overdraft task completion

The key finding from this experiment was that a significant proportion of the participant cohort (36.8%) failed to obtain an overdraft; in contrast, overall task completion rates for the menu-listed options balance and order statement were high (>90%). This leads to the conclusion that the system-initiated proposal – in its present form – is unsuitable as a method for introducing new, hidden, menu options into the dialogue of a mass-market automated telephone service.

The main reason for participants failing to obtain the overdraft appears to emanate from their procedural and declarative knowledge of the automated service (i.e. how it works and how to operate it) – commonly referred to as the user's mental model. A number of factors contribute to shaping this mental model: previous use of the service, experience of using other similar applications, information obtained from user guides, knowledge about how speech recognition technology works and so on. The user's mental model may not always be consistent with the system's conceptual model (Wærn 1993), an issue observable in the current experiment with the concept of hidden menu options. The de-briefing interview revealed that the dominant interpretation of how to operate the automated service was to select an option from the main menu: participants did not say "overdraft" because there was no such option in the menu to choose from. Furthermore, the navigational route length indicated that participants did not simply hang up once they had determined that the overdraft option was not included in the main menu: most participants looped through the main menu more than once before giving up. Essentially, these participants knew what they had to say, and where in the dialogue they should say it, but were prevented from doing so due to their failing to extrapolate beyond their ascribed strategy of selecting options from the menu. In fact, their conviction was so strong that they did not even consider attempting to just say "overdraft". These findings suggest that hidden menu options introduce dissonance in the user's mental model of the service.

This finding about the user's mental model of the service is interesting, but hardly surprising, considering that the concept of menu selection is enforced in most of today's automated telephone services. A closer examination of the navigational path length for participants (*N* = 72) who managed to complete the overdraft task further suggests that the concept of menu selection featured more prominently overall and that the hidden menu option therefore caused confusion: 23 of these participants (31.9%) had at least one extra turn-taking iteration (silence or chose an alternative option) after having listened to all options of the two main menu halves. What, then, enabled some participants to succeed with the overdraft request in the dialogue where others failed? The experiment results suggest that there were two major factors at play: volunteering behaviour and gender differences.

The first of these two – volunteering behaviour – derived from participants' propensity to say "order statement" during the first half of the main menu (before they had heard the option listed). The experiment results showed that participants who volunteered the statement keyword also were more successful at completing the overdraft task, which in turn suggests that they were less likely to abide by the main menu listings. The reason behind this kind of volunteering behaviour is not clear, but the results indicate that their approach may have been 'accidental' rather than strategic in that they may not have been fully aware that they did not always wait for the option to be listed in the main menu before selecting it; this is supported by the finding that six participants who volunteered "statement" in at least one of their calls did not volunteer the "overdraft" keyword. One possible explanation behind this behaviour is that these participants' responses were triggered by the reading task instructions presented on the priming sheet, rather than by the menu options in the system prompts. Based on the experiment findings, it is not possible to determine whether or not the volunteering behaviour would feature in the same way in real-world use of the automated service.

The second contributing factor to overdraft success/failure in the experiment concerns user gender differences: the fact that, unexpectedly, significantly more male than female participants managed to complete the overdraft task. The impact of gender difference on overdraft task success was particularly noticeable when taking into account statement volunteering behaviour: the link between volunteering behaviour and overdraft task completion was stronger in the female participant group than in the male. Specifically, females who never volunteered "statement" had a low rate of completing the overdraft task, in comparison with females who did volunteer "statement" and both the volunteering and the non-volunteering males. In other words, male participants in the non-volunteering group seemed to be able to, or more willing to, try to adopt different strategies (rather than selecting from the main menu) in order to complete the overdraft task. These findings are interesting and suggest that male and female users may have different abilities or traits, or that they may adopt different problem solving strategies, when using automated telephone services. Psychometric analyses were outside the scope of the current research and therefore the underlying factors responsible for this behaviour could not be identified. Future studies that aim to explore the issues with hidden menu options should aim to include psychometric tests to assess participants' abilities.

5.6.5 The impact of overdraft task completion

To test the impact of overdraft task completion (success/failure) this factor was brought into the analysis of the differential scores (UQ2-UQ0) as a covariate. Unsurprisingly, results showed that failing to complete the overdraft task in the fourth phone call had a significantly negative impact on user attitudes toward service usability, both overall and for individual statements in the questionnaire. Log files and comments from the exit interview supported these findings by revealing how participants struggled to get to terms with the concept of asking for a service option which was not present in the main menu. The conclusion is that, for these kind of mass-market self-service menu-driven applications, 'hidden menu options' are not a viable strategy.

5.6.6 The impact of eligibility for overdraft

One consequence of offering an overdraft to all callers in the Welcome stage of the dialogue is that ineligible customers may be encouraged to apply for an overdraft and, subsequently, have to be turned down. This is likely to have a negative impact on user attitudes toward the service. Comparing differential scores for eligible/ineligible participants (UQ2-UQ0) there were signs that having to turn down applicants could have a negative impact on perceived service usability (based on the limited sample size $N = 28$ available for these analyses, $p = .056$).

Furthermore, although not recognised as a problem by the participants, a SID offer in the Welcome stage of the dialogue could become irritating with repeat use of the automated service – especially if it is an offer that only applies or appeals to a limited set of service users. Considering the indifference in participant responses (differential scores and SIDQ), it is concluded that the Welcome location does not offer any tangible advantages for a SID delivery and that the ID&V and Transaction locations are preferable.

In sum, in accordance with experiment prediction 3, having to turn down ineligible callers (who experience the offer at the Welcome location) had a significantly negative impact on user attitudes to the service.

Chapter 6

Be not too tame neither, but let your own discretion be your tutor: suit the action to the word, the word to the action.

- William Shakespeare (1564-1616) -

Experiment 3 – Dialogue delivery strategies for system-initiated digressive proposals

6.1 Introduction

Experiment 1 identified two SID offer delivery strategies: Signpost and Follow-on. Participants who experienced the Signpost strategy were more positive toward receiving an overdraft offer compared with the Follow-on group. It was not established (based on the evaluation of application usability) what it was in the Signpost offer that might have triggered a more positive response. A complementary questionnaire (SIDQ), with focus on the SID offer itself, was devised for Experiment 2 to broaden the usability evaluation. The current study revisits the issue with using contrasting SID delivery strategies and uses the SIDQ to provide a more detailed evaluation.

The concept of 'hidden' menu items caused confusion among participants in Experiment 2: some 40% of participants failed to successfully complete an overdraft request. A perhaps obvious solution to this problem would be to make the availability of the overdraft option more transparent to the caller by simply adding 'overdraft' to the existing main menu listing. The overdraft was included in the main menu listing in the current experiment and the successfulness of this approach was evaluated by instructing participants to use the automated service to apply for an overdraft facility on their account (task completion).

6.2 Design objectives

Two design objectives have been introduced and will be explored in the current experiment: evaluation of two contrasting SID delivery strategies (Signpost and Follow-on) and task success rates when the overdraft option is included in the main menu. The rest of this section is concerned with a more detailed description of the design consideration and the implementation particulars involved.

6.2.1 Dialogue engineering objective 1: Strategy for delivery

The current experiment employed the two SID delivery strategies developed in Experiment 1: Signpost and Follow-on. Both SID offers were located immediately following a balance request (this is the Transaction SID location proposed in Experiment 2).

6.2.2 Dialogue engineering objective 2: Prompt register

The wording of the Signpost offer was as follows:

> ***"You might like to know that you can have an overdraft on your current account. To find out more, just say overdraft at the menu of services".***

The contrasting Follow-on offer delivery strategy began with a short message and then engaged the customer in a "yes/no" response dialogue which allowed interested customers to say "yes" and pursue the offer immediately (system prompts for error-level repeats are provided in Appendix 4.1):

> ***"You might like to know that you can have an overdraft on your current account. Would you like to arrange an overdraft now?"***

Customers who declined the offer were given a Signpost message with information about how to go about applying for an overdraft:

> ***"If you would like to apply for an overdraft in the future, just say overdraft at the menu of services."***

6.2.3 Contributing factors: overdraft completion

Participants were able to access the overdraft application dialogue by selecting "overdraft" at the menu of services. An important result found in Experiment 2 was that participants failed to ask for an overdraft at the menu when the overdraft word was not included in the list of service options. In Experiment 2 the proposal message contained instructions to "just say overdraft" but the menu of services did not have an explicit overdraft option to select – it was up to the caller to initiate the request. Results showed that some 40% of participants did not say "overdraft", despite being heavily primed to speak their commands and having a task sheet at hand instructing them to apply for an overdraft. This result indicates that users of self-service telephone applications rely on selecting the option from a menu and that "just say overdraft at the menu of services" is not a strong enough cue. In order to solve this problem in the SID dialogue considered here, the overdraft option was added to the second menu level (Figure 4); updated prompts are shown in Table 36. The overdraft option was also assigned a DTMF option, consistent with the current design of the automated banking service.

MAIN_MENU_A	error=0	-no change-
	error=1	-no change-
	error=2	You can choose from balance, recent transactions, overdraft, funds transfer, item search, order statement or change TIN. Please say the name of the service you would like or say help for further details.
MAIN_MENU_B	error=0	In addition you can select overdraft, funds transfer, item search, order statement, or change TIN. Which service would you like?
	error=1	Please say overdraft, funds transfer, item search, order statement or change TIN.
	error=2	You can choose from balance, recent transactions, overdraft, funds transfer, item search, order statement or change TIN. Please say the name of the service you would like or say help for further details.
HELP	first time help requested	At this point you can get the balance on your account by saying balance, hear a list of the latest transactions by saying recent transactions, apply for an overdraft by saying overdraft, transfer money between your own accounts by saying funds transfer, search for a specific item on your account by saying item search, request an account statement through the post by saying order statement or change your secret telephone identification number by saying change TIN. Please select one of these options or say help for further details.
	second time help requested	If you would like to use your telephone keypad, for balance, press 1; for funds transfer, press 3; for order statement, press 4; for item search, press 5; for recent transactions, press 6; for overdraft, press 7; for change TIN, press 8. Which service would you like?

Table 36. Updated prompt recordings used in the main menu dialogue (including new overdraft option).

Previous experiments featured simplified overdraft application dialogues in which the applicant could simply accept or reject an amount proposed by the automated service. A more elaborate overdraft application dialogue was implemented for Experiment 3 which enabled the applicant to specify an amount. Two versions of the overdraft application dialogue were implemented: one version included the maximum overdraft limit allowed on the account the system prompt upfront (the 'shadow limit') whereas the other version did not reveal this information (the overdraft negotiation dialogue is outside the scope of the current research, however, detailed descriptions of flow-charts and prompts are provided in Appendix 4.1).

6.3 Experiment predictions

The primary aim of the experiment was to assess participants' attitudes to the SID strategy. The main experiment predictions were as follows:

1. Based on the findings from Experiment 1, it was predicted that neither the presence of a SID (overdraft offer) nor delivery strategy would have a significant impact on participants' attitudes towards the usability of the automated service.

2. The questionnaire (SIDQ), developed in Experiment 2, would be used in the current research to explore participants' attitudes towards the SID offer strategy. It was predicted that the Follow-on strategy would be perceived more negatively in terms of length and intrusiveness compared with the (shorter) Signpost strategy.

3. Experiment 2 revealed that a significant number of participants did not succeed in locating the hidden overdraft option in the main menu. The current experiment aimed to address this problem by including an overdraft option in the menu of services. It was predicted that this would assist callers in completing an overdraft request and that task completion would reach the same levels as for 'balance' and 'order statement' in Experiments 1 and 2.

6.4 Method

Chapter 3 provided an overview of the experiment method adopted in the current research. This section provides further details that are relevant and specific to SID Experiment 3.

6.4.1 Design

Experiment analyses rely mainly on a between-group design; repeated-measures, within-group comparisons were indirectly achieved by running between-group analyses on the change in attitude – the differential scores – computed by subtracting baseline questionnaire scores (obtained after two practice phone calls) from questionnaire scores obtained after the third (SID offer). Grouping (between-group) variables included age, gender and offer strategy (Signpost and Follow-on).

6.4.2 Participants

A cohort of 179 participants (88 males and 91 females) contributed to the evaluation in Experiment 3[31]. Participants were recruited from the general public and only 11 of them had previous experience of using PhoneBank *Express*.

6.4.3 Materials

Participants were given the following personal banking details (described further in section 3.3): a membership number, a TIN, details of two accounts (one savings account and one current account). As in Experiment 2, participants were not given any priming materials or pamphlets on how to operate the automated service; instead they were told that they would be using an automated telephone banking service – PhoneBank *Express* – and they could speak their commands or use the buttons on the telephone keypad. No other instructions on how to use the service (i.e. which buttons to press or which words to use) were given.

6.4.4 Procedure

Participants were assigned to one of the three experiment conditions at random (No-proposal control group, Signpost and Follow-on). Upon arrival, the participant was greeted and asked to take a seat by the telephone. The participant received instructions about the experiment and was then given a sheet containing the fictitious persona details (see Section 3.7 for a more detailed description). The participants' tasks were to make telephone calls (five in total) to the automated banking service to find out the balance of 'their' current account and then order a printed statement for 'their' savings account. Participants were asked to take a note of the balance (the

[31] A total of 18 participant data sets have been excluded from the analysis as they did not complete all phone calls to the automated service, or because they had not experienced the SID offer.

balance fluctuated slightly in order to keep participants' engagement in monitoring the accounts). These two tasks were then repeated through each of the first three phone calls. On the third call participants were exposed to an overdraft proposal and in the ensuing phone calls the participant's task was to apply for an overdraft on the current account (fourth call) and then to modify an existing overdraft amount (fifth call). The participants were instructed to write down the result of each overdraft application. All participants took part in the overdraft application phone calls in the experiment; control group participants, who did not experienced a SID offer, were told that "the automated service now enables you to apply for an overdraft by saying *overdraft* at the menu of services" before making their fourth phone call to the service.

The findings from the fourth and fifth phone calls (usability evaluation) were important in order to develop suitable negotiation strategies for the overdraft application dialogue. However, this additional segment of the research is not directly related to issues surrounding the design of digressive dialogues and has therefore, in the interest of brevity, been excluded from the results and conclusion sections in this thesis. Task completion rates from the fourth phone call (first overdraft application task) will be presented and discussed.

The experiment session proceeded in a number of clearly defined stages which are outlined in Table 37 below. All participants in the experiment made their first two phone calls to the same, core, version of the automated service (without SID offers). Following the completion of the second call, participants were then asked to complete an attitude questionnaire (Appendix 1.4) to establish the reference level of the usability of the service; this questionnaire will be referred to here as 'UQ0'.

During the third call to the service, participants (except those in the control group) experienced the overdraft proposal following the current account balance enquiry. After this phone call participants completed the same attitude questionnaire but focussing on their experience of the service in last call (referred to as 'UQ1'). Additionally, after establishing that the participant had noticed the information about the overdraft, the 'SIDQ' was administered (for participants in the Signpost and

Follow-on groups) with the instructions that the use of 'overdraft proposal' in the questionnaire referred to the overdraft information experienced.

Experiment stage	**Experiment condition**	**Materials used**
Welcome, introduction, priming	Same for all participants	Persona details
Two phone calls to core service	Same for all participants	Task sheets (in each call obtain balance; order printed statement)
Usability assessment	Same for all participants	Usability questionnaire (UQ0)
One phone call to service with proposal **3 versions implemented**	1: No-proposal control group	Task sheet (obtain balance; order printed statement)
	2: Signpost SID strategy	
	3: Follow-on SID strategy	
Usability assessment		Usability questionnaire (UQ1) SID offer questionnaire (SIDQ)
Two phone calls to service **2 versions implemented**	1: Informative overdraft application dialogue	Task sheets: In call 1: setup overdraft In call 2: modify existing overdraft limit
	2: Non-informative overdraft application dialogue	
De-briefing interview		De-briefing interview Demographic questionnaire

Table 37. Overview of Experiment 3 procedure.

In the fourth phone call, participants were instructed to contact the service to apply for an overdraft (followed by a fifth call in which they were instructed to modify the amount) and note down if they perceived that they had managed to complete the overdraft task; attitudes to the usability of the overdraft application dialogue were captured after each of these phone calls. The experiment session was then ended with a de-briefing interview and the demographics/technographics questionnaire. A summary of the experiment design is provided·in Table 38.

Title	Experiment 3: strategy for delivery	
Design		One independent sample, between-subjects design adopted
Predictions	E3.1	The system-initiated digression would have no impact on participant attitudes to service usability.
	E3.2	The SID strategy would have little impact on participants attitude to service usability. In the analysis of delivery strategy itself: the Follow-on SID would be perceived to be more intrusive.
	E3.3	Adding the overdraft service option to the main menu listing would result in majority of participants completing an overdraft request.
Independent variables	1	Application: service version (3 levels)
	2	Participant: gender (2 levels)
	3	Participant: age group (3 levels)
Dependent variables		Usability questionnaire, 'UQ0' and 'UQ1' (1-7 Likert scale) SID questionnaire, 'SIDQ' (1-7 Likert scale)
Other data		De-briefing interview
Location		University Research Centre, central Edinburgh
Participant cohort		*N* = 180 (target, 30 participants in each experiment condition, six treatment permutations: 3 service versions and 2 overdraft application dialogue versions)
Remuneration		£20
Duration		Approximately 50 minutes

Table 38. Summary table of the SID strategy Experiment 3.

6.5 Results

The results analysis presented in this section was based on data entries from participants (*N* = 179) who had managed to successfully complete all their phone calls to the service. Results include: demographic/technographic details, task completion rates, service usability, SID evaluation ratings and de-briefing interview data.

6.5.1 Demographic/technographic data

Table 39 details the participant age and gender distribution for each experiment condition. The sample was overall well balanced by gender, although some cells were slightly over represented. There was also a bias evident towards the youngest age group, consequential from the recruitment process. About 37.4% of participants (*N* = 67) stated that they had used an automated telephone banking for their personal banking needs, prior to taking part in the experiment.

		Experiment condition			
Age group	Gender	No-proposal control group	Signpost SID strategy	Follow-on SID strategy	Total
18-35 years	Male	19	21	22	62
	Female	22	18	18	58
36-49 years	Male	4	2	3	9
	Female	5	9	5	19
50+ years	Male	8	7	2	17
	Female	4	3	7	14
Total		62	60	57	*N* = 179

Table 39. Analysis of participant cohort by age, gender and SID strategy.

6.5.2 Task completion

System log data obtained in Experiments 1 and 2 showed that overall task completion rates for both the balance request and statement order were high (>90%) and similar levels were achieved in the current experiment. In terms of overdraft completion rates, all participants managed to request an overdraft in the fourth phone call (setting up an overdraft) and only one participant failed to request an overdraft in the fifth phone call (modifying an existing overdraft).

6.5.3 Usability ratings prior to experiencing the SID (UQ0)

In the third phone call to the service, participants experienced one of two system-initiated overdraft offers (there was also a no-offer control group who just used the standard version of the service). Participants' attitudes towards the service were measured both immediately prior to experiencing the SID, following their second practice phone call (UQ0), and then again after completing the phone call with the SID delivery (UQ1). Responses to the usability questionnaires were analysed, both in terms of overall mean scores and according to means for individual attributes (per statement analysis).

A univariate ANOVA was run on participant responses from UQ0 with age and gender as between-group variables (Table 40). There were no significant differences overall.

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Between-participant effects					
AGE	3.803	2	1.901	2.293	.104
GENDER	.064	1	.064	.077	.781
AGE * GENDER	1.099	2	.550	.663	.517
Error	143.458	173	.829		

Table 40. ANOVA on overall usability mean scores (UQ0).

Univariate ANOVAs (with between-group variables age and gender) were also performed on each of the individual 20 UQ0 statements; score profiles for main factors are shown in Chart 26 and Chart 27.

Some individual questionnaire items were statistically significant with regards to age (Chart 26): *feeling flustered* when using the service [*df* = 2, *F* = 7.651, *p* = .001], this time the youngest age group rated the service significantly more negatively compared with both the mid- and older age groups [*p* = .003 and *p* = .021 respectively]; *feeling under stress* [*df* = 2, *F* = 3.828, *p* = .024] with the youngest age group being significantly more negative than the mid age group [*p* = .008].

Furthermore, in terms of *knowing what to do* (competency) [*df* = 2, *F* = 5.457, *p* = .005] and e*ase of use* [*df* = 2, *F* = 4.002, *p* = .020], Post Hoc tests revealed statistically significant differences [*p* = .009 and *p* = .036 respectively] between the youngest and oldest age groups. There was only one significant effect found for gender (Chart 27): females found the service more *polite* compared to male participants [*df* = 1, *F* = 4.261, *p* = .040].

There were no significant interactions between factors age and gender.

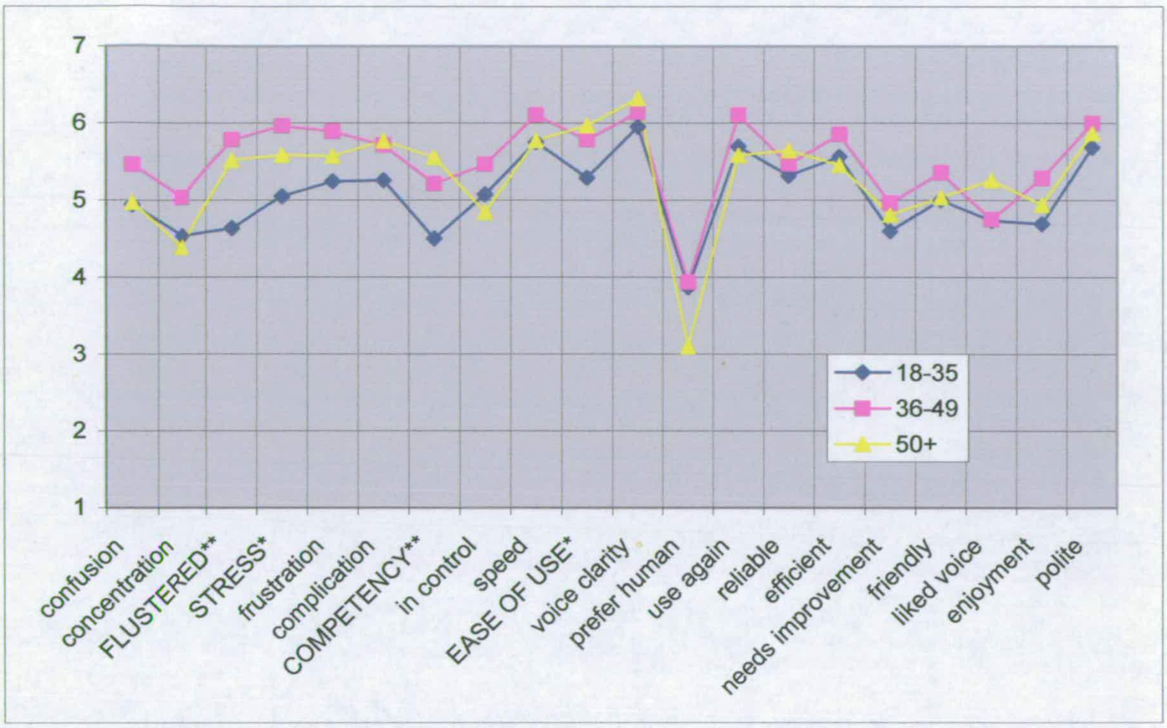


Chart 26. Main scores for UQ0 attributes, split according to age factor with three levels [p*<.05; ***p*<.01].**

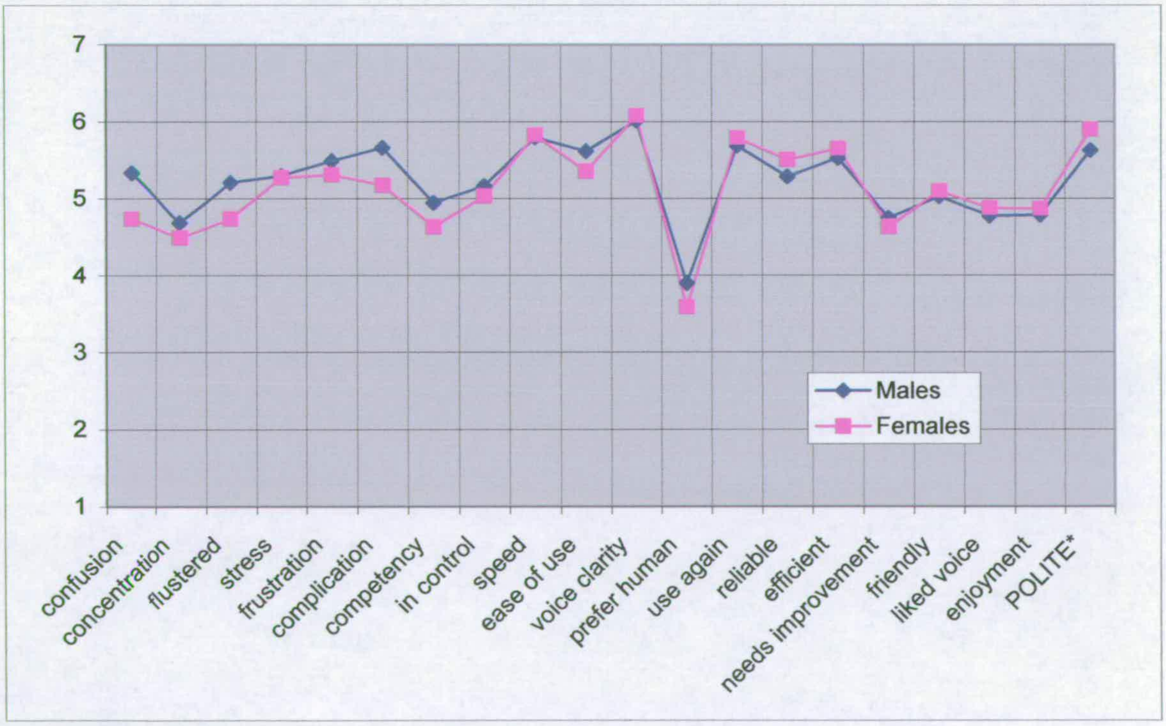


Chart 27. Main scores for UQ0 attributes, split according to gender [p*<.05].**

6.5.4 Changes in usability ratings following SID (UQ1-UQ0)

The second set of analyses concerned the impact of the presence of a SID on participants' attitudes towards service usability. This change in attitude was measured using the differential scores, computed by subtracting the UQ0 scores from the UQ1 scores. A univariate ANOVA was then run with age, gender and offer delivery (two levels: SID present and SID absent, referred to as SID_Y/N) as between-group variables (Table 41). There were no statistically significant differences for main effects of age, gender, offer delivery or Intercept value leading to the conclusion that presence of a proposal did not have an impact on participants' attitudes towards service usability overall.

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Between-participant effects					
Intercept	.168	1	.168	.647	.422
AGE	.796	2	.398	1.535	.219
GENDER	.231	1	.231	.892	.346
SID_Y/N	.132	1	.132	.509	.477
AGE * GENDER	.429	2	.215	.828	.439
AGE * SID_Y/N	.014	2	.007	.028	.972
GENDER * SID_Y/N	.543	1	.543	2.096	.150
AGE * GENDER * SID_Y/N	.577	2	.288	1.113	.331
Error	43.293	167	.259		

Table 41. Univariate ANOVA on differential usability mean scores (UQ1-UQ0), all participants (*N* = 179).

Differential attitude scores (UQ1-UQ0) were also computed for each of the individual questionnaire statements and univariate ANOVAs (with the same between-group variables) were performed. Differential score profiles for factors age, gender, and SID presence are shown in , Chart 29 and Chart 30 respectively.

The Intercept values for each of the statements reflect the change in attitude for the participant group as a whole (irrespective of the presence/absence of a SID) compared with '0' (no change). Results showed a positive increase (*M*=0.6384) in participants' feeling that they *knew what to do* [*df* = 1, *F* = 5.361, *p* = .022] and *liking the voice* [*M* = 0.2179, *df* = 1, *F* = 11.469, *p* = .001]. There had, however, been a decline (*M* = -0.1173) in attitudes to the *clarity of the voice* [*df* = 1, *F* = 6.617, *p* = .011].

Only a few items were statistically significant for main effect in the analysis of individual questionnaire statements. There were moderate significant main effects for age (Chart 28) with regards to finding the service *easy to use* [*df* = 2, *F* = 4.767, *p* = .010]; Post Hoc tests revealed that it was the difference between the youngest and oldest age group that was statistically significant [*p* = .011]. In the gender analysis (Chart 29), female participants were significantly less happy about *using the service again* [*df* = 1, *F* = 10.399, *p* = .002]. Participants who experienced the version of the automated service without a SID offer (Chart 30) found it *more confusing to use* than did participants who were subjected to an overdraft offer [*df* = 1, *F* = 4.670, *p* = .032]. This last finding (that the control group found the service more confusing) appears counter-intuitive; however, with just one such significant (lower level of significance) item among 20, this is likely to have occurred by chance.

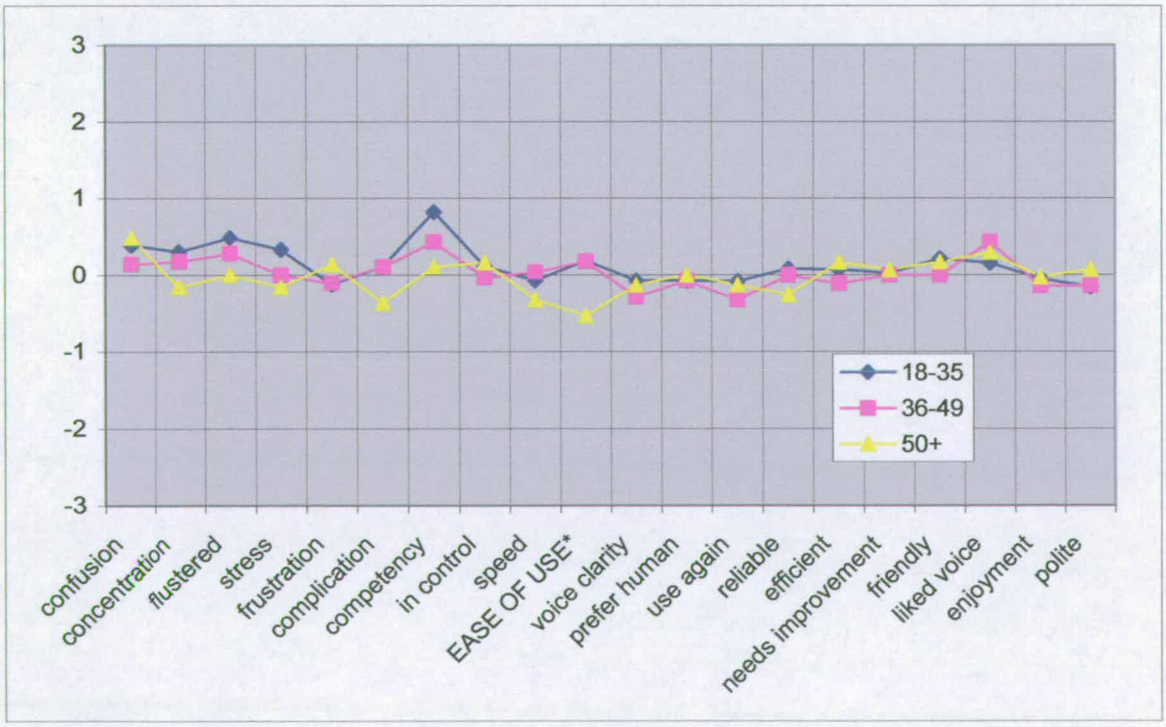


Chart 28. Differential score profiles (UQ1-UQ0), split according to age group [*p<.05], all participants (N = 179).

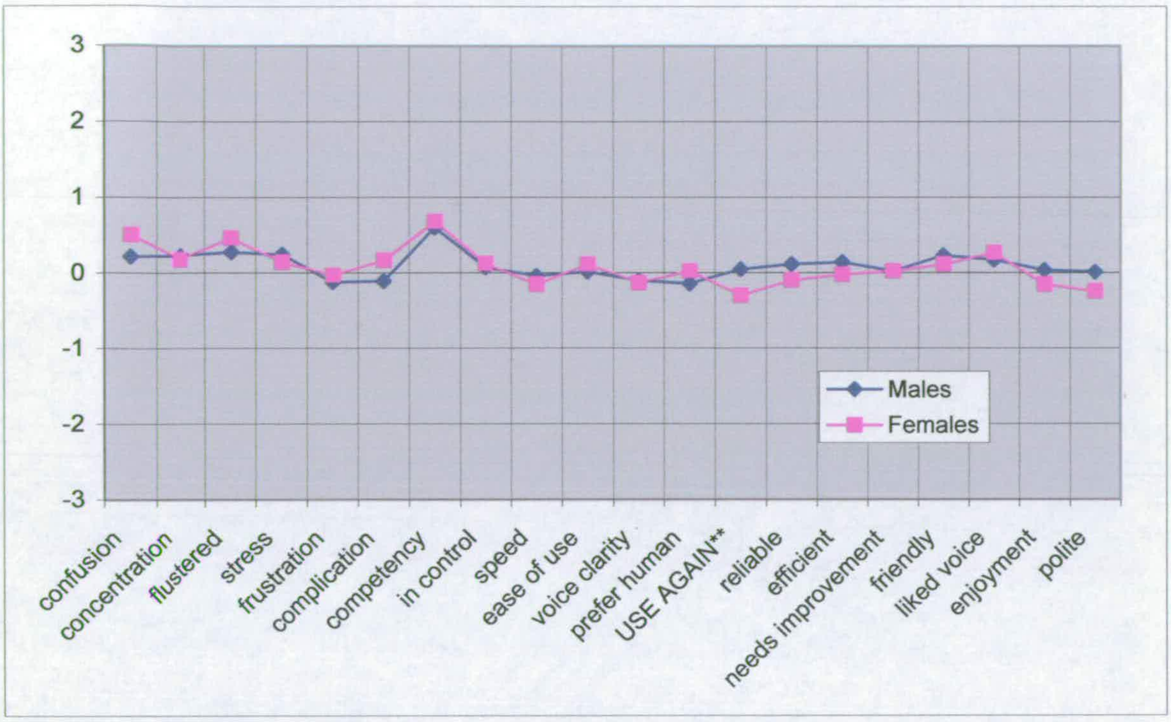


Chart 29. Differential score profiles (UQ1-UQ0), split according to gender [$p<.01$], all participants ($N = 179$).**

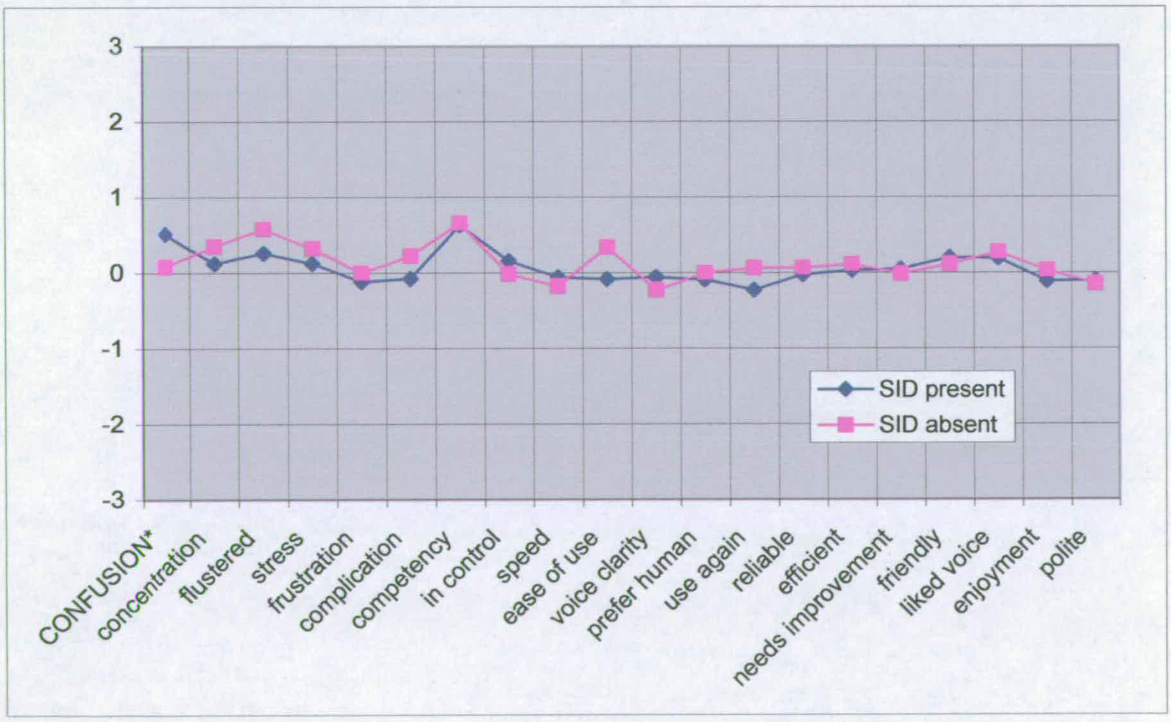


Chart 30. Differential score profiles (UQ1-UQ0), split according to presence/absence of SID offer [*$p<.05$], all participants ($N = 179$).

There were also a number of interactions that were moderately significant in the analysis. These are summarised in Table 42.

Questionnaire item	Interaction	Statistics
I felt flustered when using the service.	GENDER*SID_Y/N	*p*=.039
I would be happy to use the service again.	AGE*GENDER	*p*=.040
I think the service needs a lot of improvements.	AGE*SID_Y/N	*p*=.035
I liked the voice.	AGE*SID_Y/N	*p*=.050
I thought the service was polite.	AGE*SID_Y/N	*p*=.014

Table 42. Summary of significant interactions for UQ1-UQ0 usability attributes, all participants (*N* = 179).

Finally, to further explore the difference in change of attitudes between the three SID strategy groups, a univariate ANOVA was run on the differential scores from participants (*N* = 117) who had experienced a SID proposal during their use of the service. Between-participant variables were age, gender and SID strategy (two levels: signpost/follow-on). The results on the overall differential scores are shown in Table 43 below.

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Between-participant effects					
Intercept	.008	1	.008	.041	.841
AGE	.837	2	.418	2.001	.140
GENDER	.006	1	.006	.027	.870
SID_STRATEGY	.163	1	.163	.781	.379
AGE * GENDER	.174	2	.087	.417	.660
AGE * SID_STRATEGY	.965	2	.482	2.308	.104
GENDER * SID_STRATEGY	.113	1	.113	.542	.463
AGE * GENDER * SID_STRATEGY	.208	2	.104	.496	.610
Error	21.949	105	.209		

Table 43. Univariate ANOVA on differential usability mean scores (UQ1-UQ0), for participants who experienced a SID offer (*N* = 117).

There were no statistically significant differences for either main effects or for interaction between variables. The *Intercept* value in Table 43 represents the overall change in attitude (positive or negative) from the value 0 which represents no change in attitude. The result obtained in the analysis indicated that the presence of a SID offer did not have an impact on overall participant attitudes [p = .841].

Univariate ANOVAs were also run on the differential scores for individual questionnaire attributes. The Intercept values for individual questionnaire items are represented by the line marked as 'SID present' in the profile Chart 30 above. Three Intercept values showed significance: after experiencing a proposal participants were more positive with respect of *knowing what to do* when operating the service [df = 1, F = 4.475, p = .037] but were *less happy about using the service again* [df = 1, F = 7.954, p = .006] and would *prefer to speak to human being*[df = 1, F = 3.938, p = .050].

In the analysis of main effect of age (Chart 31), two items were statistically significant: *flustered* [df = 2, F = 3.221, p = .044] and *under stress* [df = 2, F = 3.221, p = .044]. Post Hoc test revealed that it was the difference between the oldest and the youngest age groups that was significant, the oldest age group feeling more flustered [p = .017] and more stressed [p = .008].

Two items were moderately significant in the analysis of questionnaire items based on gender (Chart 32); female participants were *less happy about using the service again* [df = 1, F = 4.733, p = .032] but were more positive about *liking the voice* [df = 1, F = 4.226, p = .042] compared to male participants. There were no statistically significant results based on SID strategy (Chart 33).

There were three items with interaction between factors; these are summarised in Table 44.

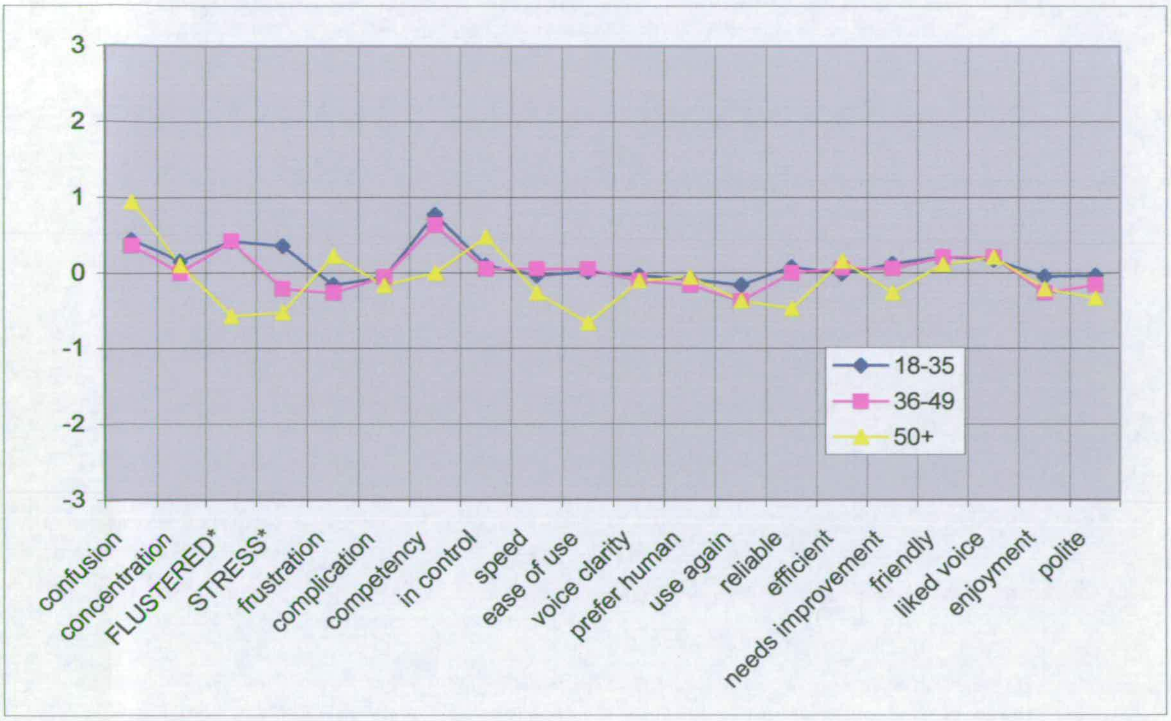


Chart 31. Differential score profiles (UQ1-UQ0), split according to age [p*<.05], for participants who experienced a SID offer (*N* = 117).**

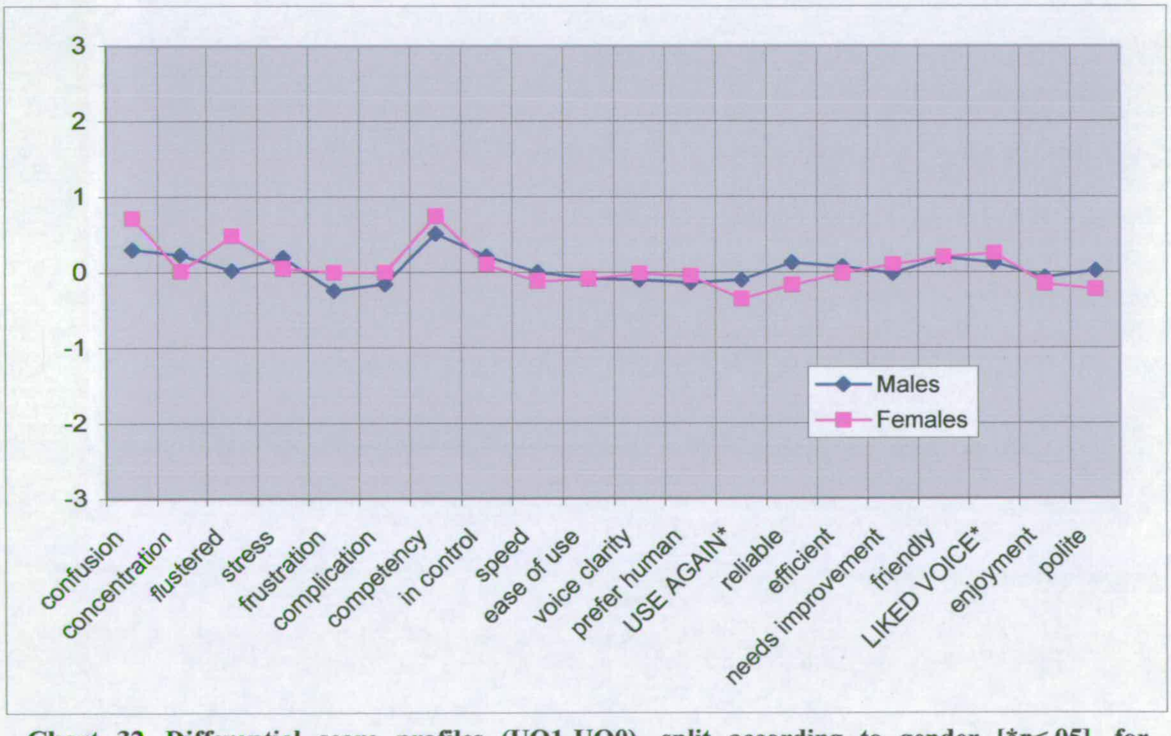


Chart 32. Differential score profiles (UQ1-UQ0), split according to gender [p*<.05], for participants who experienced a SID offer (*N* = 117).**

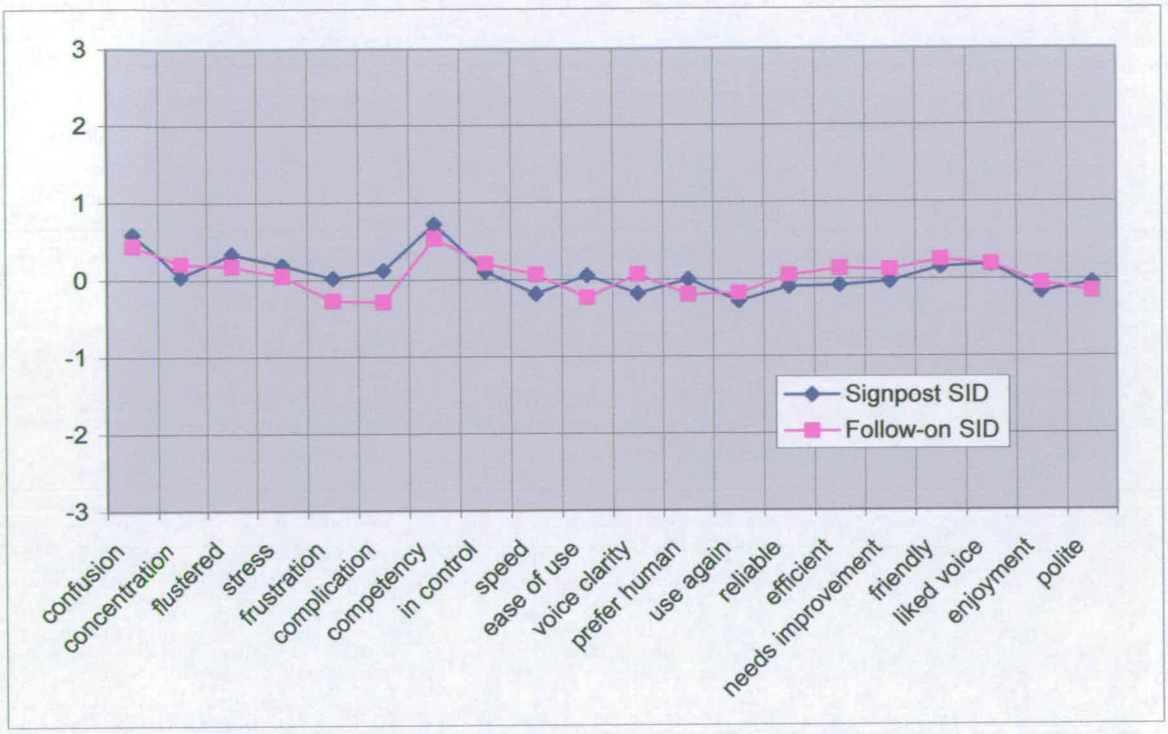


Chart 33. Differential score profiles (UQ1-UQ0), split according to SID strategy, for participants who experienced a SID offer (*N* = 117).

Questionnaire item	Interaction	statistics
I felt flustered when using the service.	GENDER*SID_STRATEGY	*p*=.040
I thought the service was easy to use.	AGE*GENDER	*p*=.043
I had to concentrate hard when using the service.	AGE*SID_STRATEGY	*p*=.013
	GENDER*SID_STRATEGY	*p*=.047

Table 44. Summary of significant interactions for UQ1-UQ0 usability attributes, for participants who had experienced a SID offer (*N* = 117).

6.5.5 Attitudes towards the SID dialogue component (SIDQ)

A univariate ANOVA was run on the SIDQ mean scores with age, gender and SID offer strategy as between-group variables (Table 45). There were no significant differences overall.

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Between-participant effects					
AGE	.839	2	.420	.412	.663
GENDER	.242	1	.242	.238	.627
SID_STRATEGY	1.366	1	1.366	1.341	.249
AGE * GENDER	1.523	2	.761	.748	.476
AGE * SID_ STRATEGY	2.580	2	1.290	1.267	.286
GENDER * SID_ STRATEGY	.548	1	.548	.538	.465
AGE * GENDER * SID_ STRATEGY	2.298	2	1.149	1.129	.327
Error	106.901	105	1.018		

Table 45. Univariate ANOVA on usability mean scores (SIDQ).

Univariate ANOVAs (with between-group variables age, gender and SID strategy) were also performed on each of the individual SIDQ statements; score profiles for main factors are shown in Chart 34, Chart 35 and Chart 36. There were no statistically significant results with age as main effect (Chart 34). Only one item was significant for gender: female participants would more strongly *prefer an overdraft proposal to be made by a human* [*df* = 1, *F* = 4.707, *p* = .032] compared to males (Chart 35).

In the analysis based on SID strategy (Chart 36), there were four significant items and it was the Follow-on participant group that took a more negative attitude to the overdraft offer; they found the offer *lengthier* [*df* = 1, *F* = 5.621, *p* = .020], that it *interrupted the call more* [*df* = 1, *F* = 6.840, *p* = .010], was *less appropriate for the service* [*df* = 1, *F* = 4.125, *p* = .045] than did the Signpost group participants. However, the Follow-on participant group felt that they *knew better how to use the service* [*df* = 1, *F* = 5.388, *p* = .022] to apply for an overdraft. Overall, it can be summarised that the Signpost SID strategy generated more positive responses in terms of perceived social characteristics (items *intrusiveness, annoyance, distraction*) and channel suitability (*interrupted the call, length, appropriateness* and *efficiency*).

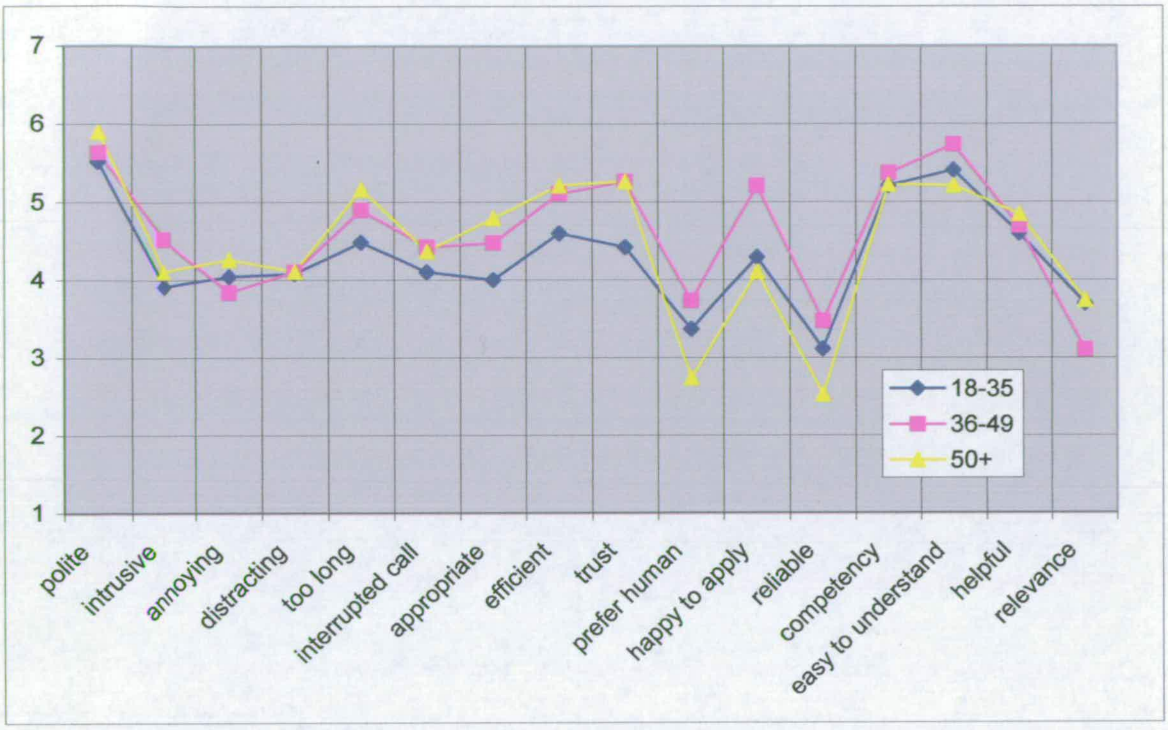


Chart 34. Mean score profiles (SIDQ), split according to age group factor, for participants who experienced a SID offer ($N = 117$).

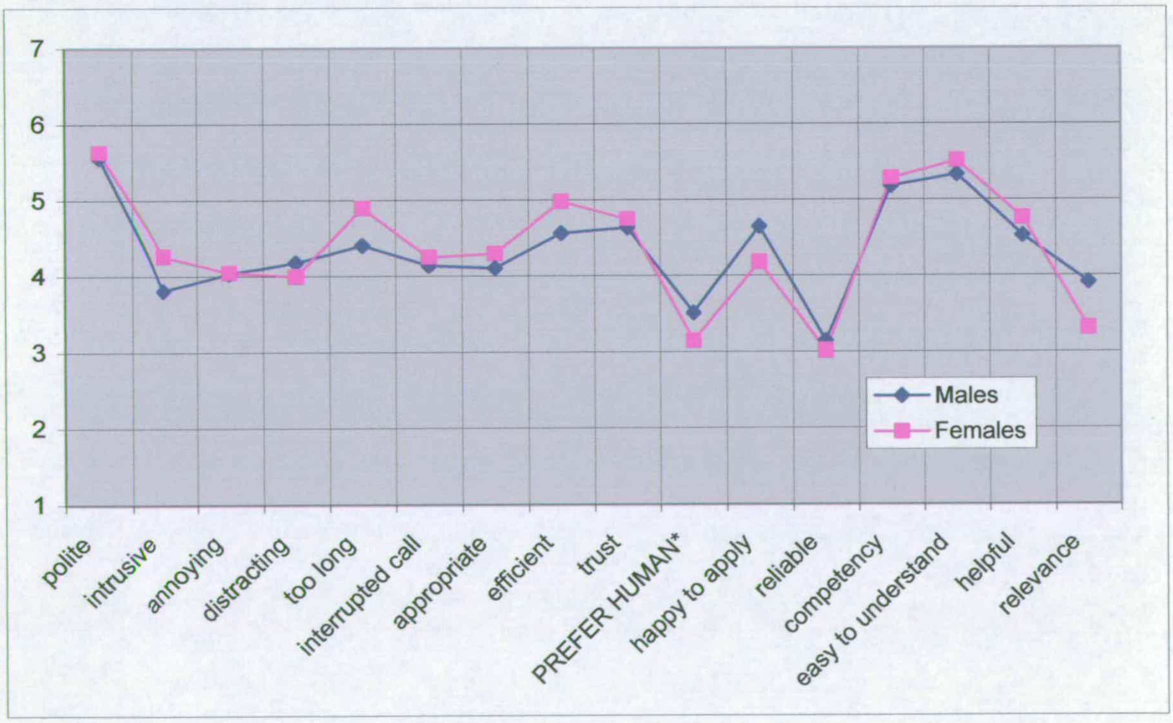


Chart 35. Mean score profiles (SIDQ), split according to gender [$^*p<.05$], for participants who experienced a SID offer ($N = 117$).

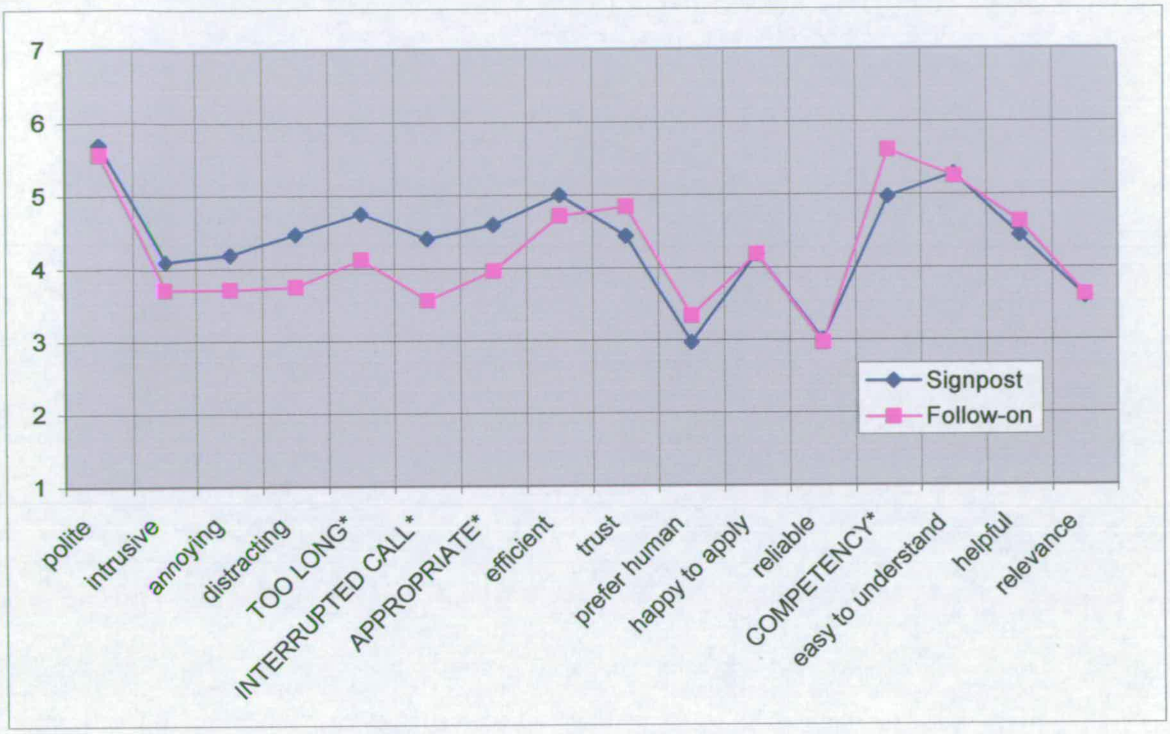


Chart 36. Mean score profiles (SIDQ), split according to proposal location [*p<.05].

Cronbach's Alpha (see Section 3.10.5) for the SIDQ in the current experiment was .89, which suggests a satisfactory level of consistency.

6.5.6 De-briefing interview feedback

All participants took part in a structured interview after their experience with the automated service. When asked if, using the automated service as a real service, they would prefer never to have been offered an overdraft 34.2% (N = 40) of participants answered "yes". Excluding the "don't know" participant responses, the preference not to hear the SID offer in the individual strategy groups was: 35.6% (N = 21) in the Signpost group and 34.5% (N = 19) in the Follow-on group (non-significant).

6.6 Conclusions

The research in this chapter centred around three themes:

1) presence/absence of SID,

2) contrasting SID strategies and

3) overdraft task completion when an overdraft option was included in the menu of services.

6.6.1 The impact of presence of SID

The prediction for this experiment was that neither the presence nor the delivery strategy of the SID would have any significant impact on user attitudes toward the usability of the service. The results from the differential scores analyses (UQ1-UQ0) confirm this prediction: there were no significant differences in responses between participants who experienced a SID offer and those that used the automated service version without a SID offer. Only one individual statement (*confusing to use*) showed statistically significant difference.

In sum, in accordance with experiment prediction 1, the presence of an overdraft offer did not have a significant impact on user attitudes to the service.

6.6.2 The impact of SID strategy

It was predicted that varying the strategy of the SID offer would have little impact on participants' attitudes to the service. The analysis of the differential scores (UQ1-UQ0) revealed that SID strategy did not have a significant impact on user attitudes. This is consistent with the findings from the evaluation of SID strategy in Experiment 1.

The current experiment used the questionnaire (SIDQ) developed in Experiment 2 to evaluate participants' reactions to the SID dialogue itself. It was predicted that the Follow-on SID strategy, being lengthier because it engaged the caller in a 'yes/no' dialogue, would be perceived as more intrusive or distracting. Results from the SIDQ analysis support this prediction, to some degree. The Signpost strategy elicited more positive participant responses with regards to intrusiveness, annoyance, distraction, length, call interruption and appropriateness. However, only three of these questionnaire items (*length of offer*, *interruption of the call* and *appropriateness*) were statistically significant. Conversely, participants who experience the Follow-on felt that they *knew better how to use the service to apply for an overdraft.* To conclude, the SIDQ questionnaire has highlighted some contrasting design qualities in the two delivery strategies originating, most probably, in the respondents' reaction to the extra dialogue step in the Follow-on strategy.

In sum, in accordance with experiment prediction 1, varying the offer strategy of an overdraft offer did not have a significant impact on user attitudes to the service. In terms of experiment prediction 2, which stated that the Follow-on strategy would be perceived as longer and more intrusive, there is some indication from the results that this was the case, however, the findings were not highly statistically significant.

6.6.3 Overdraft task completion

In Experiment 2, where a hidden overdraft menu option was introduced, only around 60% of participants succeeded with an overdraft request. In light of the high task completion rates for options that are explicitly listed in the menu of services (i.e. balance and order statement), an obvious solution was to include the overdraft option in this listing. This was the approach adopted in the current experiment and an overdraft option was inserted in the second half of the main menu listing.

Although this strategy 'cured' the problems associated with the hidden menu option (resulting in that all participants managed to request an overdraft in phone call four) it is a counterproductive solution in many other ways. Firstly, touch-tone button options are used in the current automated service and there are limits to the number of single key options that can be assigned to additional product offerings. Longer and more complex key sequences would have to be used to accommodate an increasing number of service options, or the menu would most likely need to be re-structured. Secondly, as a consequence of using touch-tone button options, each time a new option is added to the automated service any printed material such as user guides need to be edited and re-distributed. Thirdly, adding more infrequently requested services (of which overdraft is an example) increases the length and complexity of menus without necessarily being beneficial across the customer user group as a whole. In the end, where SIDs are used to introduce new options into the dialogue of an automated telephone service, main menu listings are inadequate and alternative ways of getting callers to request service options should be sought.

In sum, in accordance with experiment prediction 3, the presence of an overdraft option in the main menu resulted in all participants successfully completing an overdraft request.

Chapter 7

Be polite to all, but intimate with few.

- Thomas Jefferson (1743-1826), US President -

Experiment 4 – The use of contrasting politeness registers for system-initiated digressive proposals

7.1 Introduction

Experiments 1-3 focussed on two dialogue engineering issues – strategy and location – for delivering system-initiated digressions in automated telephone services. It was concluded that varying the strategy and location did not have a significant impact on participants' attitudes to the usability of the service. Furthermore, when a system prompt register was adopted which tones with the remaining prompts in the core automated service, the deployment of a digression did not have a negative impact on user attitudes. These findings are reassuring and support the use of SIDs in automated telephone services.

The experiment detailed in this chapter extends these findings by exploring further the use of contrasting prompt registers in the offer. Specifically, the SIDs used here explicitly stated in the opening phrases that the offer constituted an interruption. This forthright method is likely to be perceived by users as more intrusive compared to the more low-key opening phrases "*you might like to know*" used in experiments 1-3, but may however better serve to alert users to the ensuing information by capturing their attention. The purpose of making deliberate digressive interruptions in the current experiment was to explore whether or not politeness registers for human-human interaction (as defined by Brown and Levinson 1987, described in Section 2.5) could be employed to mitigate the adverse effects of these dialogue intrusions. This was achieved by deploying Follow-on strategy SID offers (described in Chapter 4) with contrasting politeness registers in the ID&V location (described in Chapter 5) in the automated telephone banking service.

The resulting politeness registers employed in the SID offers represents particularly pronounced forms of the positive, negative and bald strategies. As such, they may be perceived to be in stark contrast with the prompting style adopted in the core service and are highly unlikely to be deployed into the dialogue flow of a real automated service, in their present form. It has already been established that SIDs that tone with the register employed in the rest of the service are well received by users. The current experiment should be considered an extension to these findings giving an opportunity to explore the explicit use of an "interruption" and how contrasting politeness strategies defined for human-human communication could be applied into human-computer interaction – an area which to date has not yet been fully explored in the domain of VUIs. For this purpose, particularly pronounced forms of politeness registers were adopted in the SIDs. To be consistent with Experiments 1-3, no changes to the core dialogue prompts were undertaken; however, for such politeness strategies to be effective, the tone and register of the whole service should be reviewed and is likely to result in new prompt designs for the core application.

The banking product selected for this SID offer was the 'Online Saver' account which offers customers preferential interest rates. Transfers to and from an Online Saver account are restricted to Internet or telephone banking (through human advisor or the automated service) making it suitable for being offered through the automated service. The SID offer was aimed at informing callers about the availability of the Online Saver account and how it could improve their savings returns.

For the purpose of the current experiment, the SIDQ questionnaire developed in Experiment 2 was modified to focus the evaluation on callers' attitudes to the contrasting politeness registers when used in the context of the automated service. The construction of the SID questionnaire is included in Section 7.4.4 below. A further FRSQ (Face Redressive Style Questionnaire) was also constructed in order to provide a qualitative analysis of the SID offer in isolation (removed from the context of the automated service) in a passive listening session at the end of the experiment procedure. The FRSQ is described in Section 7.4.5 below.

7.2 Design objectives

Three design objectives have been introduced and will be explored in the current experiment: 1) presence of a SID offer (explicit interruption), 2) impact of contrasting politeness registers when interrupting the service dialogue and 3) evaluation of the contrasting politeness registers when heard in isolation. The rest of this section is concerned with a more detailed description of the design considerations and the implementation particulars involved.

7.2.1 Dialogue engineering objective: Prompt register

The resulting three proposal variants have the following basic design criteria in common: they start out with an explicit interruption (mitigated by contrasting politeness registers); they point out the financial benefits to the customer; they give details of restrictions that apply (that transfers to and from Online Saver accounts can only be done via telephone or Internet banking); and, finally, they allow interested customers to pursue the offer immediately by engaging them in a "yes/no" ('follow-on') dialogue. For customers who answer "no" at this point the service dialogue continues with "Would you like another service?". Participants who answer "yes" to the proposal hear the following message (note that the actual application procedure was simplified for experimental purposes):

> ***"Thank you, your new Online Saver account will be available from tomorrow".***

For the purpose of the experiment, the system-initiated proposals were deployed immediately after a caller had been uniquely identified (ID&V location). The wordings for each of the three contrasting styles of proposals are detailed below (for further details on face redressive registers as proposed by Brown and Levinson, see section 2.5).

The style of speaking adopted in the prompt messages are likely to have a significant impact on how callers perceive the contents of the proposal and the underlying characteristics of the speaker. During the recording session, the voice talent was not given any specific instructions on how to read out the three contrasting SID offers; the intent being to make the offers sound as 'natural' as possible. Several versions of the proposals were recorded and the three which were selected for deployment in the dialogue were very similar in intonation and tone of voice. There were some words in the texts with some extra emphasis ('stress') and these have been underlined in the resulting prompt wordings (Table 46, Table 47 and Table 48).

Positive face-redressive prompt register

Brown & Levinson's theory states that threats to the addressee's positive face (through an interruption) are mitigated by using expressions of solidarity, informality and familiarity. Examples of positive face-redress registers are, exaggerating the interest in the addressee; sympathising with the addressee; and avoiding disagreement. In the current experiment the Positive face-redress was realised by the following linguistic devices (Brown & Levinson 1987:101-129):

Being optimistic: *"I know you won't mind..."*

Informality: *"...cutting in..."*

Intensifying interest with the addressee: *"...special information for you..."*

Exaggerating approval with addressee: *"...make your growing savings grow even more."*

Presupposing common ground: *"we all want the best return possible..."*

Showing concern for the addressee's wants: *"with your interests in mind, I suggest..."*

Offering and promising: *"...an Online Saver account that will give you better interest:..."*

Giving or asking for reasons: *"why not set one up today!?"*

Personalising speaker and addressee: *"Do you want me to do that for you now?"*

Positive face-redress – (prompt recording 30 seconds long)
"I know you won't mind me cutting in with some special information for you, about how to make your growing savings grow even more. We all want the best return possible from our savings. With your interests in mind, I suggest you open an 'Online Saver account' that will give you better interest than the accounts you've got just now. You can transfer money to and from an Online Saver account through telephone or Internet banking. Why not set one up today! Do you want me to do that for you now?"

Table 46. The Positive face-redressive SID offer. Words with emphasis (stress) are underlined.

Negative face-redressive prompt register

Negative face-redress involves expressions of restraint, formality and distancing, such as being conventionally indirect, giving deference and apologising. In the current experiment the negative face-redress was realised by the following linguistic contents (Brown & Levinson 1987: 129-211):

Apologising: *"I'm very sorry to interrupt..."*
Stating the face-threatening act as a general rule: *"it is the bank's policy to notify..."*
Impersonalising speaker and addressee: *"...notify customers how to..."*
Being indirect: *"we wish to inform you..."*
Giving deference: *"...as a valued customer..."*
Being pessimistic: *"you may therefore want to consider..."*
Going on record as not indebting addressee: *"we would be happy to..."*

Negative face-redress – (prompt recording 31 seconds long)
"I'm very sorry to interrupt, but it is the bank's policy to notify customers about how to improve their savings returns. We wish to inform you, as a valued customer, that an 'Online Saver account' offers better interest than the accounts you hold at present. You may therefore want to consider opening an account of this type. Transfers to and from Online Saver accounts are made through telephone or Internet banking. We would be happy to set up an Online Saver account for you today. Would you like us to do that now? "

Table 47. The Negative face-redressive SID offer. Words with emphasis (stress) are underlined.

Bald (no face redress) prompt register

Undertaking a speech act without positive or negative face-redress is described by Brown & Levinson (1987) as performing the act 'baldly'. In contrast to the registers used to mitigate positive and negative face threats, the primary concern in the *Bald register* is to be direct and concise. The Bald register is applied under circumstances where the face threat can be ignored, in the interest of urgency and efficiency. The speaker might, for example, feel that the information is so important or interesting to the addressee such that there is no need for a more convoluted expression. Alternatively, the speaker might be unconcerned about any imposition on behalf of the addressee. The Bald proposal in the experiment was stripped of any kinds of face-redress and started with: *"I'm interrupting to inform you about..."*. The absence of more convoluted face redressive registers makes the Bald strategy naturally shorter than both the Positive and Negative politeness styles.

Bald – (prompt recording 18 seconds long)
"I'm interrupting to inform you about how to improve your savings returns. The 'Online Saver account' offers better interest than the accounts you have at present. You can transfer money to and from an Online Saver account through telephone or Internet banking. Do you want to set up an online saver account now?"

Table 48. The Bald (no face-redress) SID offer. Words with emphasis (stress) are underlined.

7.3 Experiment predictions

The primary aim of the experiment was to assess the relative effectiveness of contrasting politeness registers when making explicit interruptions in SID offers. The main experiment predictions were as follows:

1. All three SID offer styles are, to some degree, intrusive in that they explicitly point out to the caller that the SID offer constitutes an interruption. It is predicted that all three contrasting SID offers will have a negative impact on user attitudes to the automated service.

2. The three SID proposals employed contrasting politeness registers. It was predicted that the 'Negative' face redressive style, which employs more deferential politeness (commonly associated with enterprise-customer relationship), would be rated more positively in terms of social appropriateness compared with the 'Positive' and 'Bald' styles. The 'Bald' style offer, being the shortest of the three SID offers, would be rated positively in terms of duration and intrusiveness.

3. It was predicted that (based on evaluations of the three politeness registers in the passive listening session) participants would perceive contrasting qualitative characteristics in the offers which would be in line with Brown and Levinson's theory of politeness.

7.4 Method

Chapter 3 provided an overview of the experiment method adopted in the current research. This section provides further details that are relevant and specific to SID Experiment 4.

7.4.1 Design

As indicated above, three contrasting politeness registers were explored. Experiment analyses rely on a combination of between-group and within-group design. Grouping (between-group) variables included age, gender and SID offer presence (offer or no-proposal control group). Politeness register (Positive, Negative and Bald) was used as a between-group variable in the analyses following first exposure to a SID offer and then as a within-group variable in the pooled data analysis after participants in the SID offer group had experienced all offer variants (controlled, randomised order).

The repeated-measure ANOVA in SPSS returns a significance value for 'Mauchly's test of Sphericity' which tests certain assumptions on the dependent variable data (check of the homogeneity of variance for all experiment factor effects). Whenever Mauchly's test of Sphericity returns a significant value ($p < .05$) the sphericity assumption has been violated and it is recommended to use the adjusted statistical values provided in the output for the within-subject factors (Field 2000, p334); the current analysis reports the Greenhouse-Geisser adjusted statistics (consequently, some degrees of freedom – *df* – values reported in the results sections will be fractional). In all other cases, the values for 'sphericity assumed' are reported.

7.4.2 Participants

A total of 111 participants (48 males and 63 females) contributed to the evaluation in Experiment 4. Participants were recruited both from the general public and from a database of Lloyds TSB customers. In total, 38 (34%) of the participants had previous experience of using PhoneBank *Express*.

7.4.3 Materials

Participants were given the following personal banking details (described further in section 3.3): a membership number, a TIN, details of two accounts (one savings account and one current account). Participants were not given any priming materials or pamphlets on how to operate the automated service; instead they were told that they would be using an automated telephone banking service – PhoneBank *Express* – and they could speak their commands or use the buttons on the telephone keypad. No other instructions on how to use the service (i.e. which buttons to press or which words to use) were given.

7.4.4 SID questionnaire

Experiment 2 identified 16 attributes (SIDQ) used to evaluate participants' attitudes to the SID offer dialogue. Modifications of the questionnaire were performed in order to adapt it to the focus of the current research. In particular, 12 new items were constructed in order to elicit respondents' attitudes to the wording and how the offer fitted in with the rest of the service. These attributes are organised into 'themes' (described below) according to the most prominent characteristics of each face redressive register and were then used to analyse participants' perceptions of the contrasting politeness styles.

The Positive face redress register was founded on 'social intimateness', implying rapport between the system and the user. The intention was to convey friendliness however, if conflicting with the listener's anticipated conversational behaviour, the message may be perceived as cajoling or socially intrusive. To explore participants' perception of the linguistic characteristics employed in the Positive face redressive proposal, the following questionnaire items were constructed:

The proposal was **friendly**.
The proposal made me feel I was **being manipulated**.
The proposal took my **interest into account**.
I found the proposal **patronising**.

The Negative face redress register employed revering phrases and 'conventional' apologetic behaviour, often associated with a customer-business relationship. The intention was to put the customer's needs in focus; however, the Negative face redress register used in the SID offer was much more pronounced than in the rest of the automated service. To explore this, the following items were added:

The style of the proposal was **too formal**.
The way the proposal was expressed was **too apologetic**.
The proposal expressed **care for my individual needs**.

Finally, the Bald offer register did not have any face redressive register and is the shortest of the three offer variants. It is less long-winded and less wordy than the face redressive offers which may appeal to customers. To assess this, two questionnaire items were included:

The proposal was very **long-winded**.
The proposal contained only **relevant information**.

In addition, three questionnaire items were included in order to establish how well the offer fitted in with the service and whether or not the participants felt they would have liked more product information:

The wording of the proposal **fitted in well with the rest of the service**.
I would want **more information before opening** an Online Saver account.
The proposal should give **more information** about the Online Saver account.

The complete questionnaire (24 items), as it appeared in the experiment, is included in Appendix 5.1.

7.4.5 FRS questionnaire

All participants took part in an evaluation of the three contrasting SID offers in a listening session towards the end of the experiment. The purpose of the listening session was to obtain a measure (manipulation check) of the *absolute politeness* (Leech 1983) associated with the face-threatening act in the SID offer, independent of the context of the telephone banking dialogue. The FRSQ extended the measure of the *relative politeness* of the proposal in the automated service obtained through the SID questionnaire (the resulting questionnaire is included in Appendix 5.2).

For this purpose, a questionnaire with 14 semantic differentials (described in section 3.10.5) was constructed. The questionnaire was split into two sections: the first set (Table 49) was introduced with the phrase "*Thinking about the proposal I've just heard, it was...*"; the second set (Table 50) was introduced with the phrase "*I associate the choice of wording in the proposal with someone who is...*". The predicted rating ordering of participant responses (based on the wordings in the offers and taking into account the domain of telephone banking services) to each of the semantic differentials is given in the middle column of the tables. For example, it was predicted – in terms of politeness – that the Negative face redressive style would be rated as more polite compared to the Positive and Bald strategies; there was no further predicted ordering between the Positive and Bald strategies for the politeness attribute.

Analyses of responses from the no-proposal delivery control group would provide an important measure of whether or not the contrasting prompt registers used in the SID offers carried significantly discernable information regarding politeness attributes, and whether or not these differences would be in line with Brown and Levinson's theory. The response data from the remaining participant groups would give an indication of participants' attitudes to the SID offer in isolation while taking into account that these individuals had already experienced the same offers in the context of the automated service.

	Predicted order	
polite	NEGATIVE > POSITIVE, BALD	impolite
formal	NEGATIVE, BALD > POSITIVE	informal
to the point	BALD > NEGATIVE, POSITIVE	long-winded
forthright	BALD, POSITIVE > NEGATIVE	diplomatic
sincere	NEGATIVE, BALD > POSITIVE	insincere
respectful	NEGATIVE > BALD, POSITIVE	patronising
personalised	POSITIVE, NEGATIVE > BALD	impersonal
apologetic	NEGATIVE > BALD, POSITIVE	unapologetic

Table 49. Semantic differentials used to evaluate attitudes to the SID offer in the listening session. Predicted ordering of the contrasting face redressive styles in the offers are included.

	Predicted order	
tactful	NEGATIVE > POSITIVE, BALD	tactless
timid	NEGATIVE > BALD, POSITIVE	self-confident
unsociable	BALD > NEGATIVE, POSITIVE	sociable
reliable	BALD, POSITIVE, NEGATIVE	unreliable
caring	NEGATIVE, POSITIVE > BALD	uncaring
unprofessional	BALD, POSITIVE > NEGATIVE	professional

Table 50. Semantic differentials used to evaluate attitudes to 'speaker characteristics'. Predicted orderings of the contrasting face redressive styles in the offers are included.

7.4.6 Procedure

Participants were assigned to one of the three experiment conditions at random: Positive face redressive offer, Negative face redressive offer, Bald non-face redressive offer and no-offer control group. Upon arrival, the participant was greeted and asked to take a seat by the telephone. The participant received instructions about the experiment and was then given a sheet containing the fictitious persona details (see Section 3.7 for a more detailed description).

The participant's task was to make a number of telephone calls (five in total for the SID offer groups and three in total for the control group) to the automated banking service to find out and take a note of the balance of 'their' current account[32]. In the third phone call to the service, three quarters of the participants were exposed to a savings proposal with varying politeness registers. These participants (excluding the control group) were instructed to make a further two phone calls to the service (to carry out the same task). These two phone calls allowed participants to experience the remaining two politeness variants in a controlled randomised order. Finally, the experiment included a passive listening session in which all participants heard each proposal over a pair of computer speakers.

The experiment session proceeded in a number of clearly defined stages which are outlined in Table 51 below. All participants in the experiment made their first two phone calls to the same, core, version of the automated service (without SID offers). Following the completion of the second call, participants were then asked to complete an attitude questionnaire (Appendix 1.4) to establish the reference level of the usability of the service; this questionnaire will be referred to here as 'UQ0'.

During the third call to the service, participants (except those in the control group) experienced the savings proposal. After this phone call participants completed the same attitude questionnaire but focussing on their experience of the service in last call (referred to as 'UQ1'). Additionally, after establishing that the participant had noticed the information about the Online Saver account (it turned out they all had), the 'SIDQ' was administered with the instructions that the use of 'proposal' in the questionnaire referred to the savings account information experienced.

[32] Participant fatigue is of concern in the case where the experiment session is lengthy (multiple phone calls and questionnaires) and, in order to reduce the time spent on each phone call, only one task was used. Furthermore, no usability questionnaires (UQ) were administered after phone calls four and five.

Experiment stage	Experiment condition	Materials used
Welcome, introduction, priming	Same for all participants	Persona details
Two phone calls to core service	Same for all participants	Task sheets (in each call obtain balance)
Usability assessment	Same for all participants	Usability questionnaire (UQ0)
One phone call to service with proposal **4 versions implemented**	1: Positive face redress SID	Task sheet (obtain balance)
	2: Negative face redress SID	
	3: Bald no face redress SID	
	4: No proposal control group	
Usability assessment		Usability questionnaire (UQ1) SID offer questionnaire (SIDQ0)
Two phone calls to service (SID groups only) **6 offer registers permutations represented**	a) (Positive): Negative, Bald b) (Positive): Bald, Negative c) (Negative): Positive, Bald d) (Negative): Bald, Positive e) (Bald): Positive, Negative f) (Bald): Negative, Positive	Task sheet (obtain balance)
Usability assessment		Attitude questionnaire after each phone call (SIDQ1, SIDQ2)
Passive Listening Session	(all 6 offer register order permutations represented)	
Usability assessment		Attitude questionnaire after each SID offer played (FRSQ0, FRSQ1, FRSQ2)
De-briefing interview		De-briefing interview Demographic questionnaire

Table 51. Overview of Experiment 4 procedure.

Participants (except those in the control.group who stopped after the third phone call) made another two phone calls with SID offers so that they had experienced all three politeness register variants, completing further questionnaires ('SIDQ1' and 'SIDQ2') after each phone call. All participants then took part in the listening session in which they listened to the three SID offers over a pair of computer speakers and, for each SID offer, completed a questionnaire ('FRSQ0', 'FRSQ1' and 'FRSQ2'). The session was then ended with a de-briefing interview and the demographics/technographics questionnaire. A summary of the experiment design is provided in Table 52.

Title	**Experiment 4: face redressive registers in SID interruptions**	
Design		One independent sample, between-subjects and within-subjects designs adopted
Predictions	E4.1	The system-initiated digression would have a negative impact on participant attitudes to service usability.
	E4.2	The Negative face redressive offer would be rated more positively in terms of social appropriateness; the Bald offer register would be rated more positively in terms of shorter duration.
	E4.3	Differences in face redressive style (listening session) would be in line with Brown and Levinson's politeness theory.
Independent variables	1	Application: service version (4 levels)
	2	Participant: gender (2 levels)
	3	Participant: age group (3 levels)
	4	SID presentation order (6 levels)
Dependent variables	1	Usability questionnaire, 'UQ0' and 'UQ1' (1-7 Likert scale) SID questionnaire, 'SIDQ0', 'SIDQ1' and 'SIDQ2' (1-7 Likert scale) FRS questionnaire, 'FRSQ0', 'FRSQ1' and 'FRSQ2' (1-7 Semantic differential scale)
Other data		De-briefing interview
Location		University Research Centre, central Edinburgh
Participant cohort		*N* = 120 (target, 30 participants in each experiment condition)
Remuneration		£20
Duration		Approximately 45-60 minutes

Table 52. Summary table of the SID politeness register Experiment 4.

7.5 Results

The results analysis presented in this section was based on data entries from participants (*N* = 111) who had managed to successfully complete all their phone calls to the service. Results include: demographic/technographic details, usability and SIDQ/FRSQ evaluation ratings and de-briefing interview data.

7.5.1 Demographic/technographic data

Table 53 details the participant age and gender distribution for each experiment condition. The sample was overall well balanced by gender, although some cells were slightly over represented. There was also a bias evident towards the youngest age group, consequential from the recruitment process.

About 55% of participants (N = 61) stated that they had used an automated telephone banking service for their personal banking needs, prior to taking part in the experiment.

		Experiment condition (3rd phone call)				
Age group	Gender	No-proposal control group	'Positive' face redress	'Negative' face redress	'Bald' face redress	Total
18-35 years	Male	5	6	6	6	23
	Female	7	7	7	7	28
36-49 years	Male	3	4	4	4	15
	Female	4	5	4	5	18
50+ years	Male	2	3	2	3	10
	Female	4	4	5	4	17
Total		25	29	28	29	N = 111

Table 53. Analysis of participant cohort by age, gender and experiment condition.

7.5.2 Task completion

Task completion rates described here were based on system log data and required that the participant had requested to hear the balance for the current account (regardless of then noting the correct balance on the task sheet). The core system dialogue does not change between phone calls and therefore it was expected that task completion rates would remain at the same high levels (>90%) as obtained in Experiments 1-3. Results showed that this was not the case in the third phone call. In this phone call, participants experienced the savings proposal and then had to accept or reject the offer to set up an Online Saver account. The task sheet did not include a task to set up a saver account and therefore all participants (as expected) answered "no". Following this, the automated service they asked participants "Would you like another service?" and participants were required to answer "yes" in order to proceed with their account balance enquiry. Answering "no" at this point ended the call to the service.

In the control group (Table 54), it was the same participant who failed to complete the balance task in calls 2 and 3 because the call had been transferred to an advisor in the account selection stage (see section 3.3.5 for a description of the dialogue). In the third phone call, one participant in the Positive face redress group did not realise (after hearing the SID offer) that the balance task should have been carried out and answered "no"; a further two participants in this group answered (wrongly) "no" and subsequently ending the dialogue. In the Bald proposal group, seven participants had answered (wrongly) "no".

	Call 1	Call 2	Call 3
Control group	26 (100%)	25 (96%)	25 (96%)
Positive face redress	29 (100%)	29 (100%)	26 (90%)
Negative face redress	28 (100%)	28 (100%)	28 (100%)
Bald face redress	29 (100%)	29 (100%)	22 (76%)

Table 54. Balance task completion rates in the first three phone calls to the automated service.

7.5.3 Usability ratings prior to experiencing the SID (UQ0)

In the third phone call to the service, three quarters of participants experienced a system-initiated overdraft offer. Participants' attitudes towards the service were measured both following their second practice phone call (UQ0), immediately prior to experiencing the SID, and then after completing the phone call with the SID delivery (UQ1). Responses to the usability questionnaires were analysed, both in terms of overall mean scores and according to means for individual attributes (per statement analysis).

A univariate ANOVA was run on participant responses from UQ0 with age and gender as between-group variables (Table 55). There were no significant effects overall.

Univariate ANOVAs (with between-group variables age and gender) were also performed on each of the individual 20 UQ0 statements; score profiles for main factors are shown in Chart 37 and Chart 38.

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Between-participant effects					
AGE	1.236	2	.618	1.554	.216
GENDER	.583	1	.583	1.467	.229
AGE * GENDER	.725	2	.362	.911	.405
Error	41.748	105	.398		

Table 55. ANOVA on overall usability mean scores (UQ0).

Four individual questionnaire items were statistically significant with regards to age: the *feeling of being under stress* [*df* = 2, *F* = 3.252, *p* = .043] with Post Hoc tests revealing statistically significant differences [*p* = .032] between the mid-age group and the older age group; the *feeling of being in control* [*df* = 2, *F* = 4.023, *p* = .020] with the difference between the youngest and oldest age groups statistically significant [*p* = .019]; the perceived *speed* of the service [*df* = 2, *F* = 3.385, *p* = .038] with the difference between the youngest and mid-age groups statistically significant [*p* = .026]; and *liking of the voice* [*df* = 2, *F* = 3.924, *p* = .023] where the difference between the mid- and oldest age groups was statistically significant [*p* = .015].

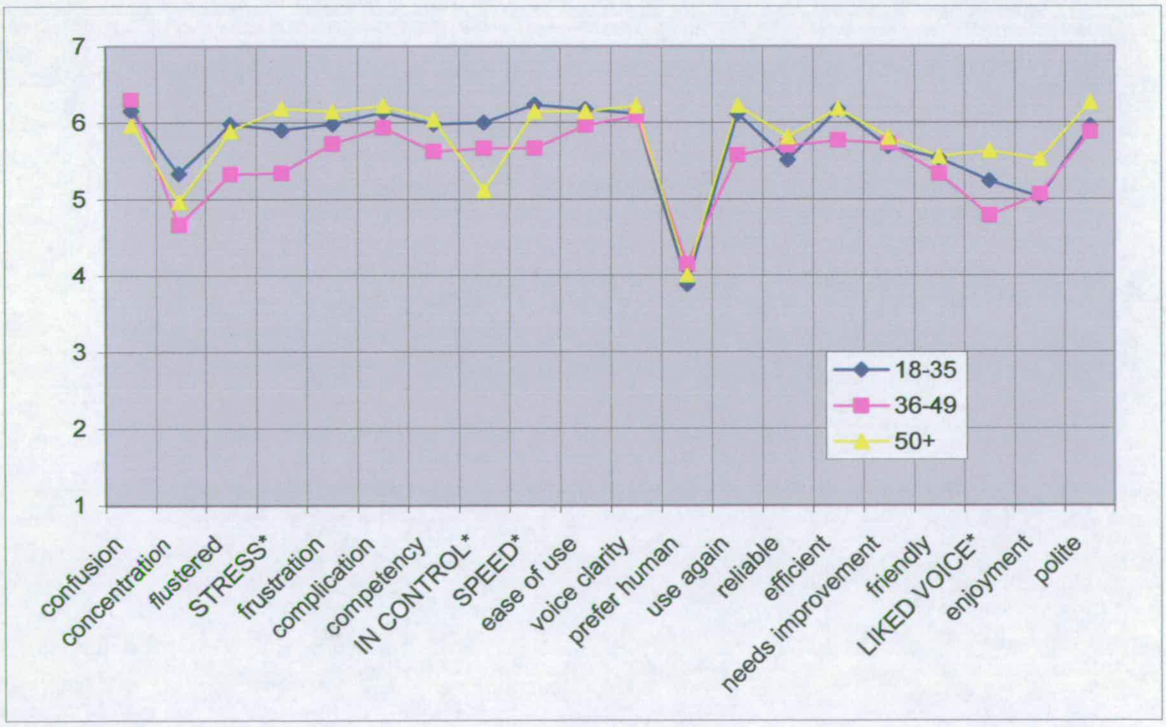


Chart 37. Main scores for UQ0 attributes, split according to age group factor with three levels [*p<.05].

There were only two significant effects found for gender (Chart 38); women were significantly more positive compared to men in terms of *knowing what to do* (competency) [*df* = 1, *F* = 5.300, *p* = .023] and *being happy about using the service again* [*df* = 1, *F* = 5.237, *p* = .024].

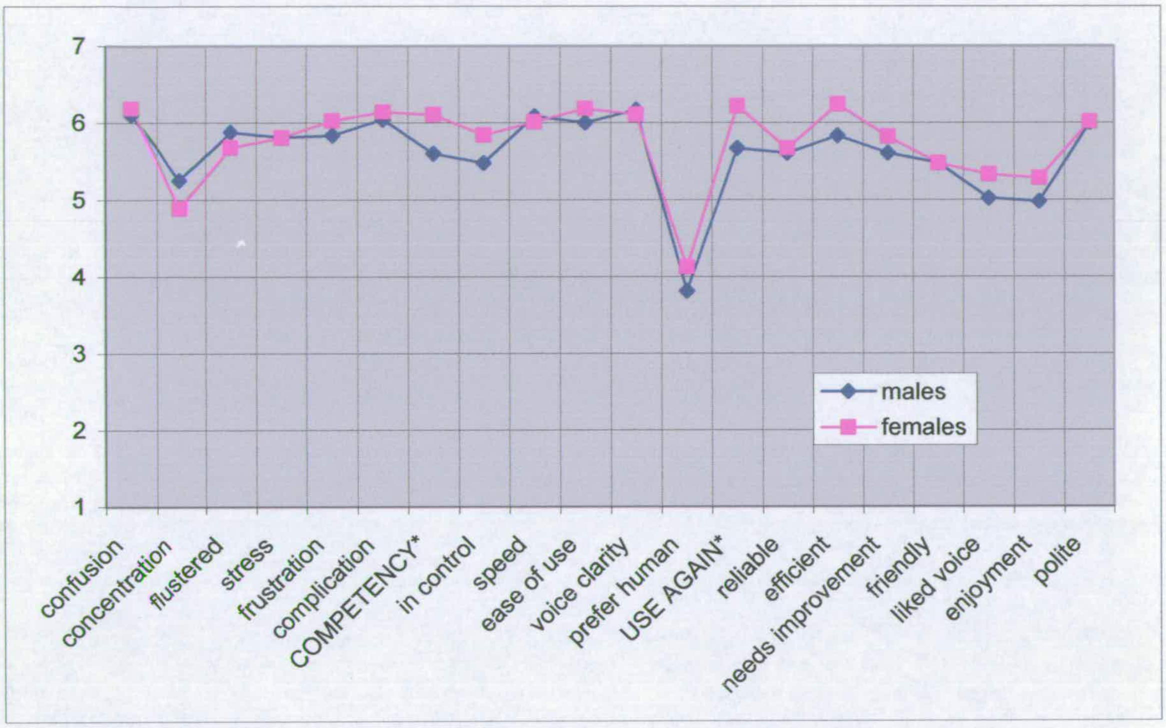


Chart 38. Main scores for UQ0 attributes, split according to gender [*p<.05].

There were no significant interactions between factors in the analyses of UQ0.

7.5.4 Changes in usability ratings following SID (UQ1-UQ0)

The second set of analyses concerned the impact of the presence of a SID on participants' attitudes towards service usability. This change in attitude was measured using the differential scores, computed by subtracting the UQ0 scores from the UQ1 scores. A univariate ANOVA was then run with age, gender and offer delivery (two levels: SID present and SID absent, referred to as SID_Y/N) as between-group variables (Table 56).

There were no statistically significant differences for main effects of age and gender. The Intercept value was highly significant indicating that overall the change in attitudes before/after the third phone call, overall, was highly significant. Furthermore, the SID_Y/N value showed that the difference in change in attitude overall between the no-proposal control group (*M* = .0069) and the SID proposal group (*M* = -.7351) was highly significant, indicative of the negative impact of the presence of SID proposals on user attitudes.

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Between-participant effects					
Intercept	9.019	1	9.019	12.059	**.001**
AGE	.111	2	.056	.074	.928
GENDER	.044	1	.044	.059	.809
SID_Y/N	11.256	1	11.256	15.049	**.000**
AGE * GENDER	.562	2	.281	.375	.688
AGE * SID_Y/N	1.498	2	.749	1.001	.371
GENDER * SID_Y/N	.234	1	.234	.312	.578
AGE * GENDER * SID_Y/N	1.643	2	.822	1.099	.337
Error	74.046	99	.748		

Table 56. Univariate ANOVA on differential usability mean scores (UQ1-UQ0), all participants (*N* = 111).

Differential attitude scores (UQ1-UQ0) were also computed for each of the individual questionnaire statements and univariate ANOVAs (with the same between-group variables) were performed. Differential score profiles for factors age, gender, and SID presence/absence are shown in Chart 39, Chart 40 and Chart 41 respectively.

There were no statistically significant differences for effects of age (Chart 39) or gender (Chart 40). Several statistically significant results were found in the analysis of questionnaire items based on presence/absence of a SID proposal (Chart 41), with participants who experienced the proposal taking a more negative attitude to the service. At the higher level of significance (p<.01), participants in the proposal group were *more stressed* when using the service [*df* = 1, *F* = 7.372, *p* = .008], felt *more frustrated* [*df* = 1, *F* = 13.592, *p* = .000], felt *less in control* [*df* = 1, *F* = 12.959, *p* = .000], were *less happy about using the service again* [*df* = 1, *F* = 11.313, *p* = .001], found the service *less efficient* [*df* = 1, *F* = 8.593, *p* = .004] and *enjoyed using the service less* [*df* = 1, *F* = 8.940, *p* = .004].

On a moderate level of significance (*p*<.05) participants found the service with the proposal to be *more confusing* [*df* = 1, *F* = 4.481, *p* = .037], required them to *concentrate harder* [*df* = 1, *F* = 4.169, *p* = .044], made them feel *more flustered* [*df* = 1, *F* = 4.087, *p* = .046], was *more complicated*[*df* = 1, *F* = 6.648, *p* = .011], they *knew less what to do* [*df* = 1, *F* = 6.041, *p* = .016], found the service *less easy to use* [*df* = 1, *F* = 4.810, *p* = .031] and felt the service *was in more need of improvement* [*df* = 1, *F* = 6.091, *p* = .015].

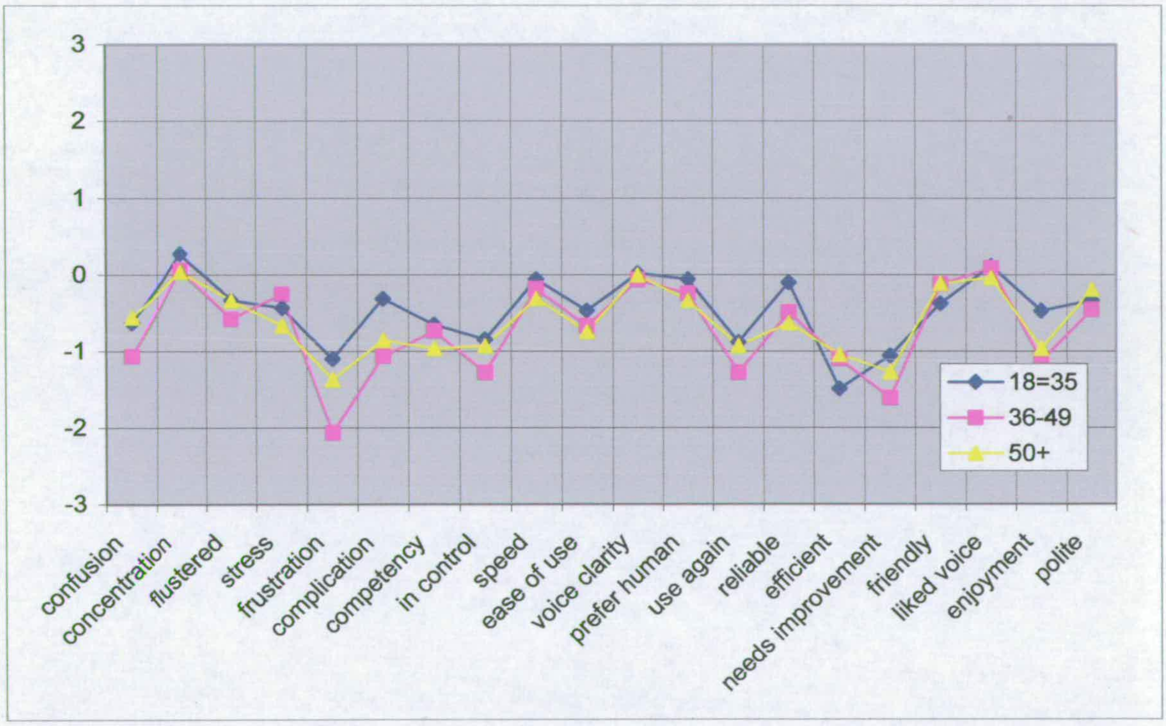


Chart 39. Differential score profiles (UQ1-UQ0), split according to age group, all participants (*N* = 111).

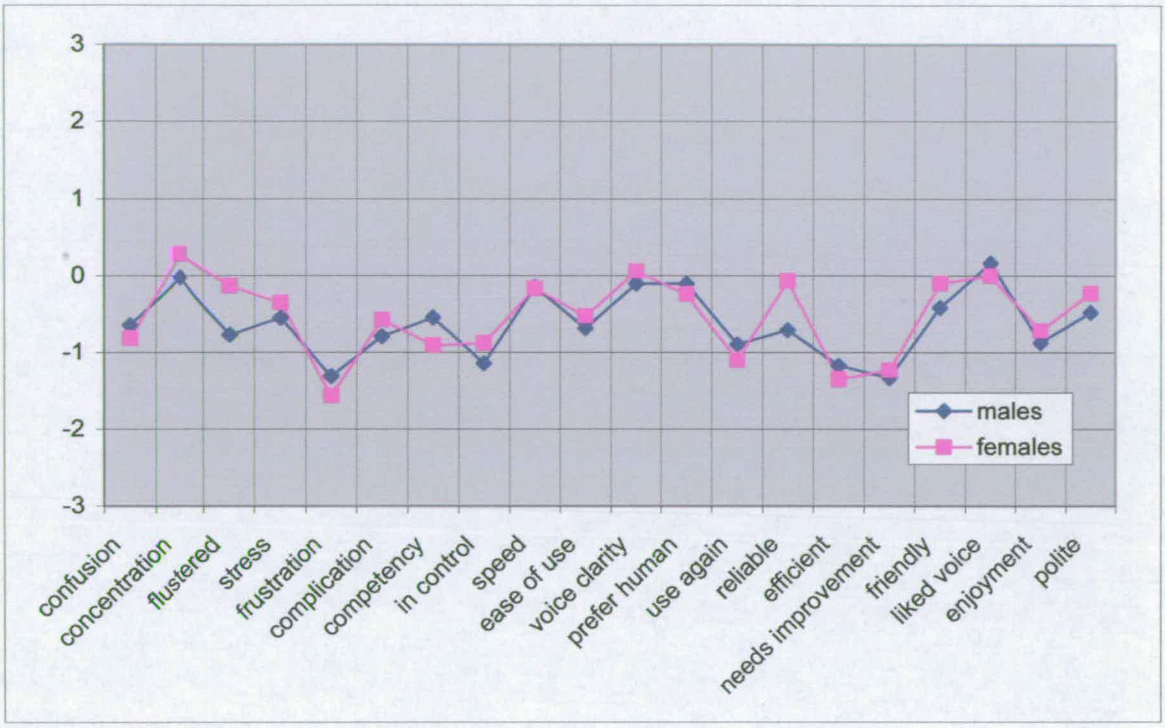


Chart 40. Differential score profiles (UQ1-UQ0), split according to gender, all participants (*N* = 111).

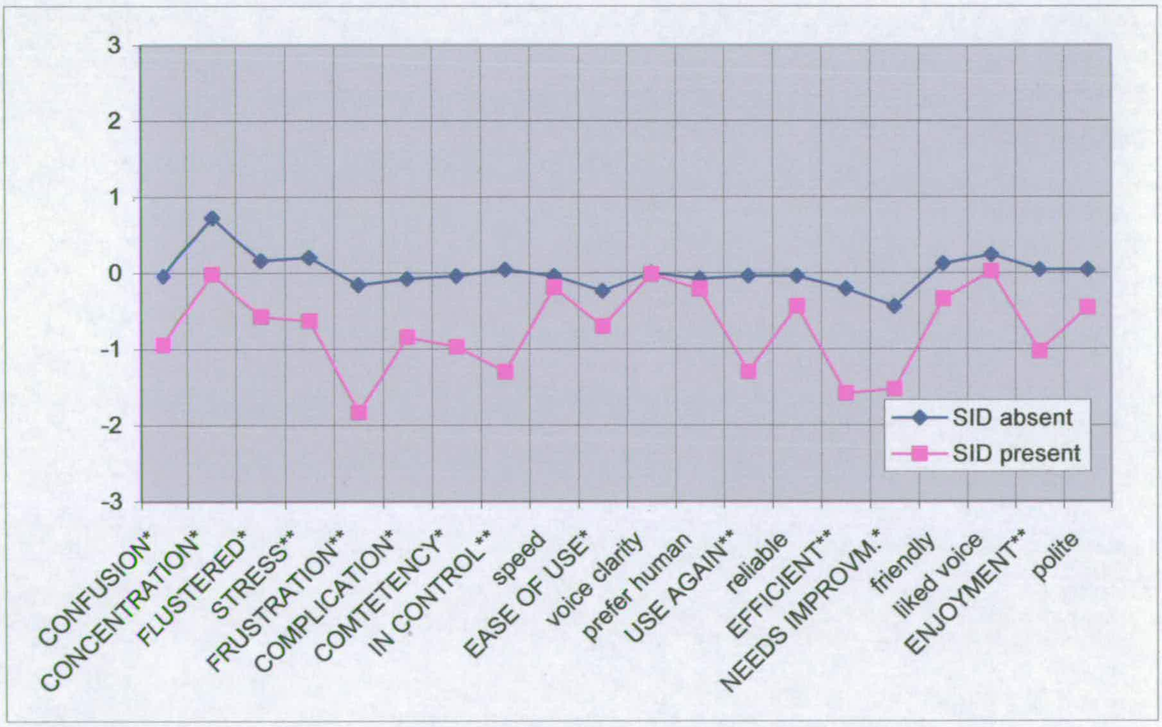


Chart 41. Differential score profiles (UQ1-UQ0), split according to presence/absence of a SID offer [*p<.05; **p<.01], all participants (*N* = 111).

There was only one two-way interaction between the factors (GENDER*SID_Y/N) for the questionnaire item ‘I found the service easy to use’ [*p* = .048]. Furthermore, there were two items with three-way interaction between the factors (AGE*GENDER*SID_Y/N): ‘I found the service confusing to use’ [*p* = .025] and ‘I felt flustered when using the service’ [*p* = .048].

Finally, to further explore the difference in change of attitudes between the three SID politeness register groups, a univariate ANOVA was run on the differential scores from the participant subgroup (*N* = 86) that had experienced a SID proposal during their use of the service. Between-participant variables were age, gender and SID register (three levels: positive, negative and bald). The results on the overall differential scores are shown in Table 57 below.

The *Intercept* value in Table 57 represents the overall change in attitude (positive or negative) from the value 0 which represents no change in attitude. The result obtained in the analysis indicated that the presence of a SID offer had a strongly significant negative impact on participant attitudes [*p* = .000]. There were no further significant main effects for factors age, gender or SID politeness register; however, there was a moderately significant interaction between factors age and SID register. Figure 21 shows the differential scores for the combined age and proposal register factors; the younger and mid-age groups responded more consistently to the SID offer (irrespective of politeness register adopted) compared to the older age group.

A one-way ANOVA on the new combined age and SID register independent variable (with nine levels) showed no overall statistical significance [*df* = 8, *F* = 1.599, *p* = .139]; and Post Hoc tests revealed no further significant differences. The only Post Hoc test which approached statistical significance was the comparison between the positive and bald registers in the oldest age group [*p* = .065].

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Between-participant effects					
Intercept	46.110	1	46.110	57.838	**.000**
AGE	2.607	2	1.304	1.635	.203
GENDER	.498	1	.498	.625	.432
SID_REGISTER	2.947	2	1.473	1.848	.165
AGE * GENDER	3.978	2	1.989	2.495	.090
AGE * SID_REGISTER	8.107	4	2.027	2.542	**.047**
GENDER * SID_REGISTER	1.511	2	.756	.948	.393
AGE * GENDER * SID_REGISTER	5.284	4	1.321	1.657	.170
Error	54.211	68	.797		

Table 57. Univariate ANOVA on overall differential usability mean scores (UQ1-UQ0), for participants who experienced a SID offer (*N* = 86); first SID offer phone call analysed here.

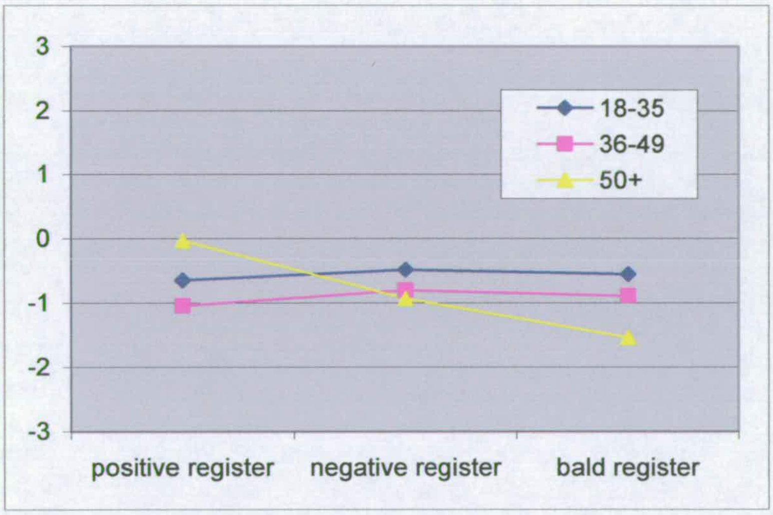


Figure 21. Overall mean usability score differences (UQ1-UQ0) grouped according to SID register and age, all SID group participants (*N* = 86).

Univariate ANOVAs were also run on the differential scores for individual questionnaire attributes. The Intercept values for individual questionnaire items are represented by the line marked as 'SID present' in the profile Chart 41. The drop in attitude for the majority of individual items showed strong statistical significance (statistical results are presented in Table 58). A further two items showed moderate statistical significance: *speed of service* [*df* = 1, *F* = 4.458, *p* = .038] and *preference of speaking to a human* [*df* = 1, *F* = 4.700, *p* = .034].

Questionnaire item	df =	F = /p =
Confusion	1	20.938/.000
Flustered	1	42.776/.000
Under stress	1	11.853/.001
Frustration	1	62.612/.000
Complicated	1	25.578/.000
Competency	1	22.582/.000
In control	1	39.012/.000
Ease of use	1	26.686/.000
Use again	1	48.364/.000
Reliable	1	62.942/.000
Efficiency	1	42.776/.000
Needs improvement	1	62.942/.000
Enjoyed using	1	46.768/.000
Polite	1	13.155/.001

Table 58. Summary of significant interactions for UQ1-UQ0 usability attributes, *Intercept values*, for participants who had experienced a SID offer (*N* = 86).

In the analysis of main effect of age (Chart 42), two items were moderately statistically significant: *frustration* [*df* = 2, *F* = 3.390, *p* = .039] and *complication* [*df* = 2, *F* = 3.083, *p* = .027]. Post Hoc tests revealed no further significant values for *frustration*; the difference in perceived *complication* between the youngest and mid-age groups, however, was significant [*p* = .039].

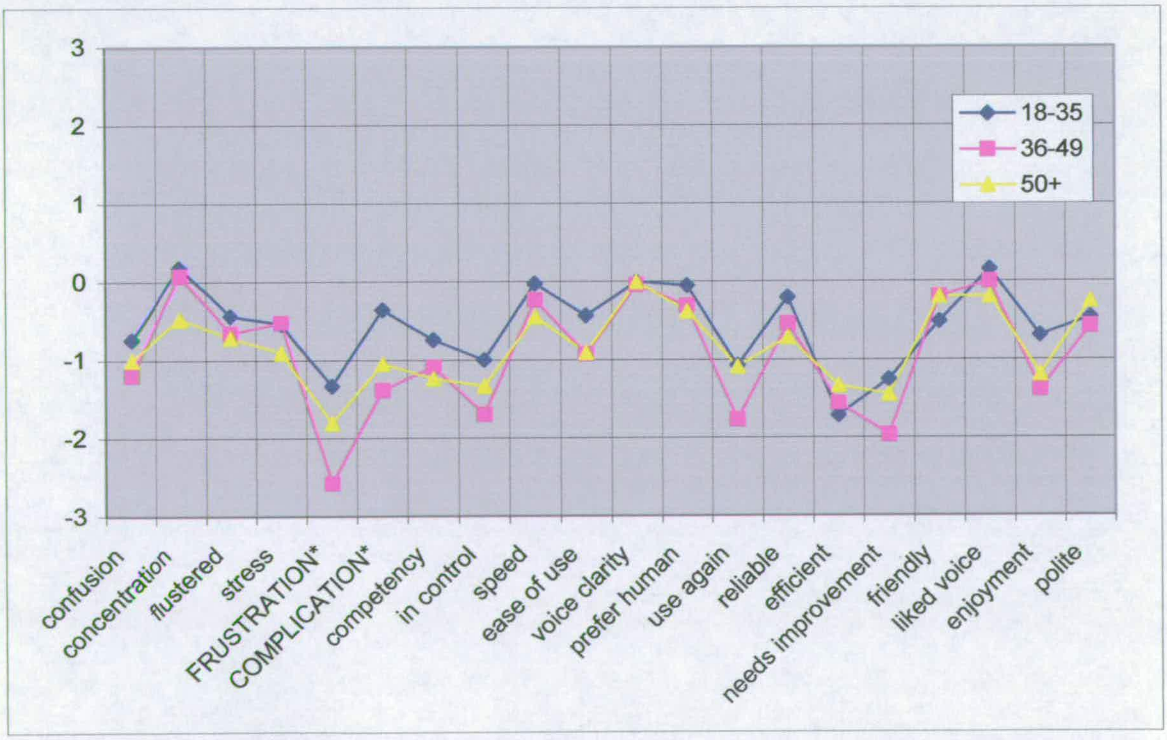


Chart 42. Differential score profiles (UQ1-UQ0), split according to age [*p<.05], for participants who experienced a SID offer (*N* = 86).

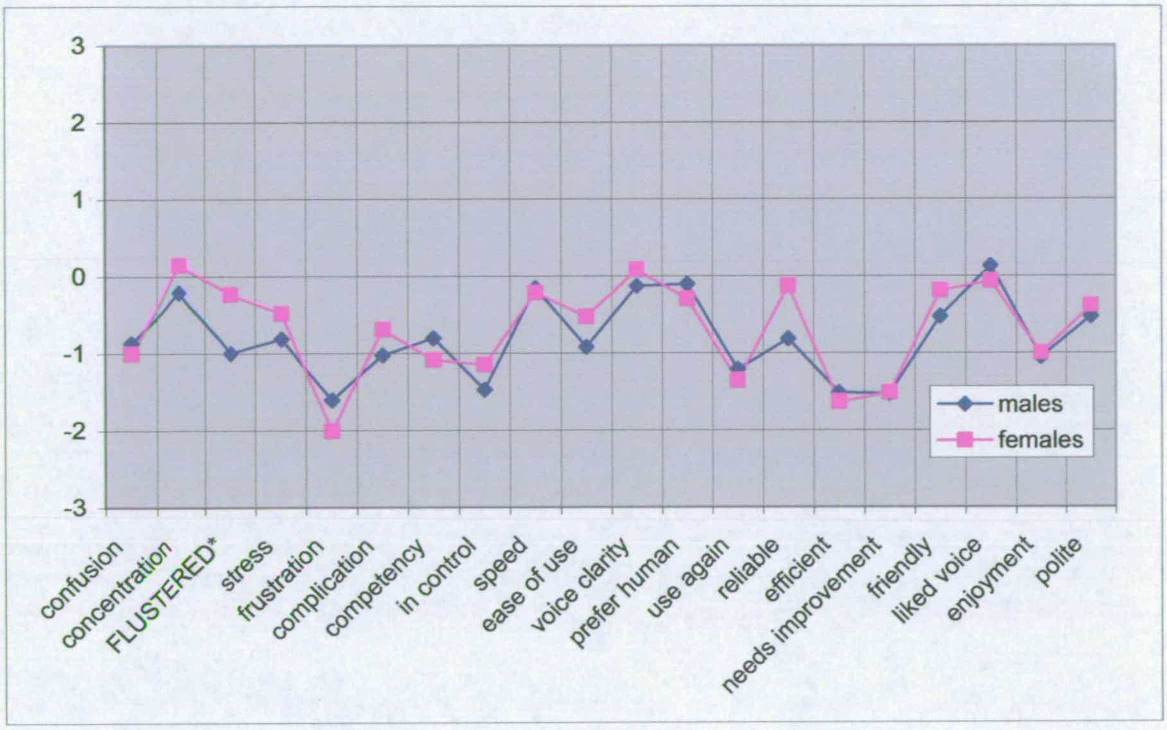


Chart 43. Differential score profiles (UQ1-UQ0), split according to gender [*p<.05], for participants who experienced a SID offer (N = 86).

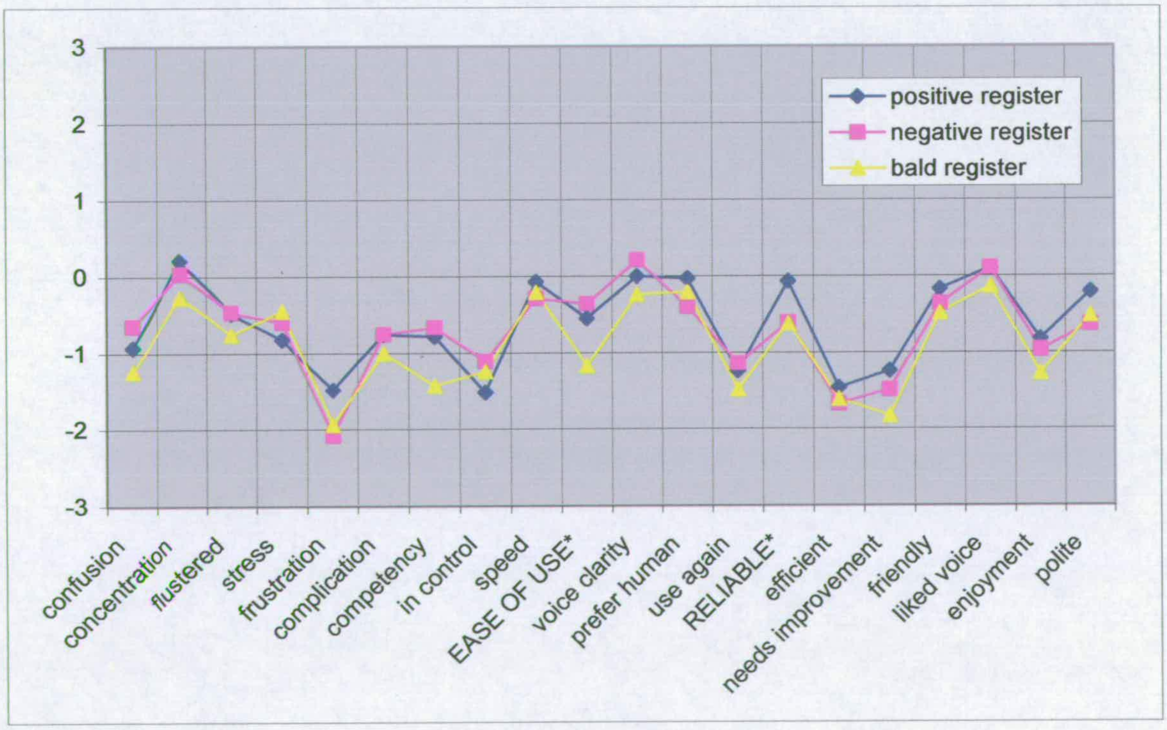


Chart 44. Differential score profiles (UQ1-UQ0), split according to SID politeness register [*p<.05], for participants who experienced a SID offer (N = 86).

Only one item was moderately significant in the analysis of questionnaire items based on gender (Chart 43); male participants felt significantly more *flustered* when using the service [$df = 1, F = 4.443, p = .039$].

Two items were moderately significant in the analysis based on SID offer register (Chart 44): *ease of use* [$df = 1, F = 4.595, p = .013$] and *reliability of the service* [$df = 1, F = 3.623, p = .032$]; Post Hoc tests revealed no further significant differences.

There were seven items with interaction between factors; these are summarised in Table 59. A further three-way interaction (AGE*GENDER*SID_REGISTER) occurred for items regarding *feeling in control* [$p = .042$], the *speed of the service* [$p = .039$], *enjoyment of using the service* [$p = .034$].

Questionnaire item	Interaction	Statistics
I found the service frustrating to use.	AGE*GENDER	p=.018
I thought the service complicated.	AGE*GENDER	p=.038
The service was too fast for me.	AGE*SID_REGISTER	p=.005
I found the service easy to use.	AGE*SID_REGISTER	p=.018
I would be happy to use the service again.	AGE*SID_REGISTER	p=.018
I felt that the service was reliable.	AGE*SID_REGISTER	p=.001
I thought the service was efficient.	AGE*GENDER	p=.020

Table 59. Summary of significant interactions for UQ1-UQ0 usability attributes, for participants who had experienced a SID offer ($N = 86$).

7.5.5 Attitudes towards the SID dialogue component (SIDQ)

The current experiment included an additional set of 24 questionnaire statements – referred to here as 'SIDQ' – addressed at capturing participants' attitudes towards the SID offer. Participants in the SID offer group ($N = 86$) completed this questionnaire after it had been established that they were all aware that an offer had been played in the third phone call, and then subsequently after each following SID offer phone call.

Two analyses were carried out on the SIDQ data. Firstly, a univariate ANOVA was run on participant responses from SIDQ0 mean scores only (first exposure to SID offer) with age, gender and politeness register as between-group variables (Table 60). There were no significant main effects or interactions between factors.

Univariate ANOVAs (with between-group variables age, gender and SID politeness register) were also performed on each of the individual SIDQ0 statements; score profiles for main factors are shown in Chart 45, Chart 46 and Chart 47 respectively.

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Between-participant effects					
AGE	4.083	2	2.041	2.250	.113
GENDER	.543	1	.543	.599	.442
SID_REGISTER	2.168	2	1.084	1.195	.309
AGE * GENDER	1.740	2	.870	.959	.388
AGE * SID_ REGISTER	4.163	4	1.041	1.147	.342
GENDER * SID_ REGISTER	.289	2	.145	.159	.853
AGE * GENDER * SID_ REGISTER	3.646	4	.911	1.004	.411
Error	61.702	68	.907		

Table 60. Univariate ANOVA on usability mean scores (SIDQ0).

Seven questionnaire items were statistically significant for main effect of age (Chart 45): *appropriateness of the proposal in the service* [*df* = 2, *F* = 3.421, *p* = .038], the *ease of understanding the proposal* [*df* = 2, *F* = 3.577, *p* = .033], the *helpfulness of the proposal information* [*df* = 2, *F* = 3.811, *p* = .027], the *feeling of being manipulated* [*df* = 2, *F* = 3.188, *p* = .048], the perception that the proposal *took personal interests into account* [*df* = 2, *F* = 6.890, *p* = .002], the degree of *formality* [*df* = 2, *F* = 5.440, *p* = .006] and the feeling that *more information about the online saver should be included in the proposal* [*df* = 2, *F* = 3.315, *p* = .042].

Post Hoc tests further revealed that the difference between the youngest and mid-age groups was significant in terms of *appropriateness* [*p* = .044] and *ease of understanding the proposal* [*p* = .026], perceived *helpfulness* [*p* = .032] – the mid-age group taking a more negative attitude in terms of these attributes. The mid-age group took a significantly more negative attitude to *the proposal taking personal* *interests into account*, compared to both the youngest [*p* = .021] and the oldest [*p* = .005] age groups. The difference between the youngest and the oldest age groups for the item *the proposal was too formal* was also significant [*p* = .011].

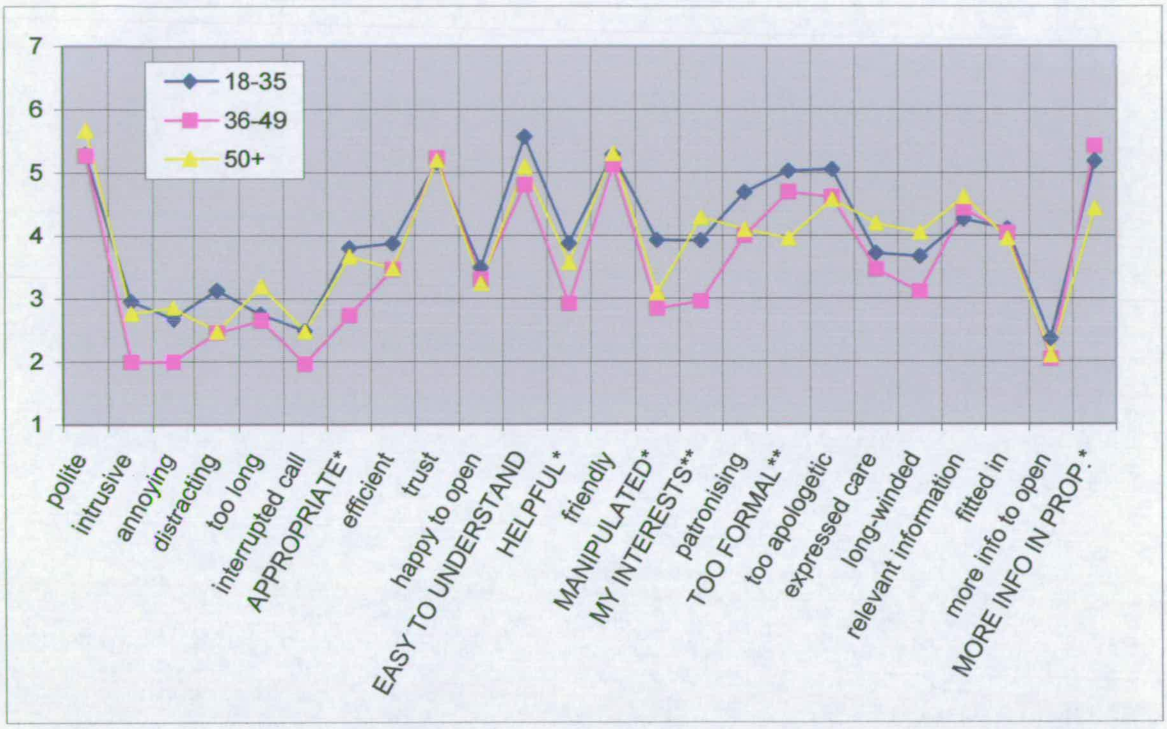


Chart 45. Mean score profiles (SIDQ0), first SID offer phone call, split according to age group factor [p*<.05; ***p*<.01], SID groups participants only (*N* = 86).**

In the analysis based on gender (Chart 46), two items were moderately significant ($p < .05$): females were more positive to the *politeness* of the proposal [*df* = 1, *F* = 4.775, *p* = .033] as well as being more positive to the *level of information* about the online saver account given in the proposal [*df* = 1, *F* = 4.226, *p* = .044].

There were no significant differences in the analysis for main effect of politeness register (Chart 47). There was one two-way interaction for AGE*SID_ REGISTER for the SIDQ item 'the proposal took my interests into account' [*df* = 4, *F* = 2.601, *p* = .044] and a further two three-way interactions (AGE*GENDER*SID_ REGISTER) for items 'the proposal was easy to understand' [*df* = 4, *F* = 2.776, *p* = .034] and 'the proposal contained only relevant information' [*df* = 4, *F* = 2.919, *p* = .027].

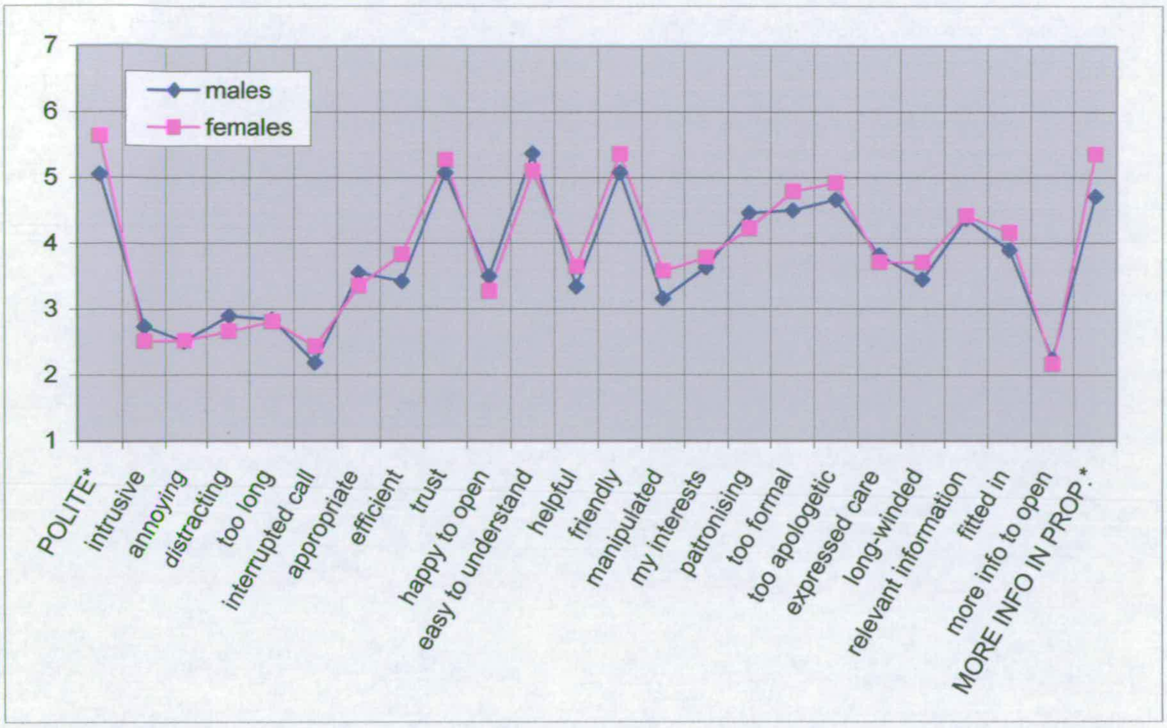


Chart 46. Mean score profiles (SIDQ0), split according to gender factor [*p<.05], SID groups participants only (N = 86).

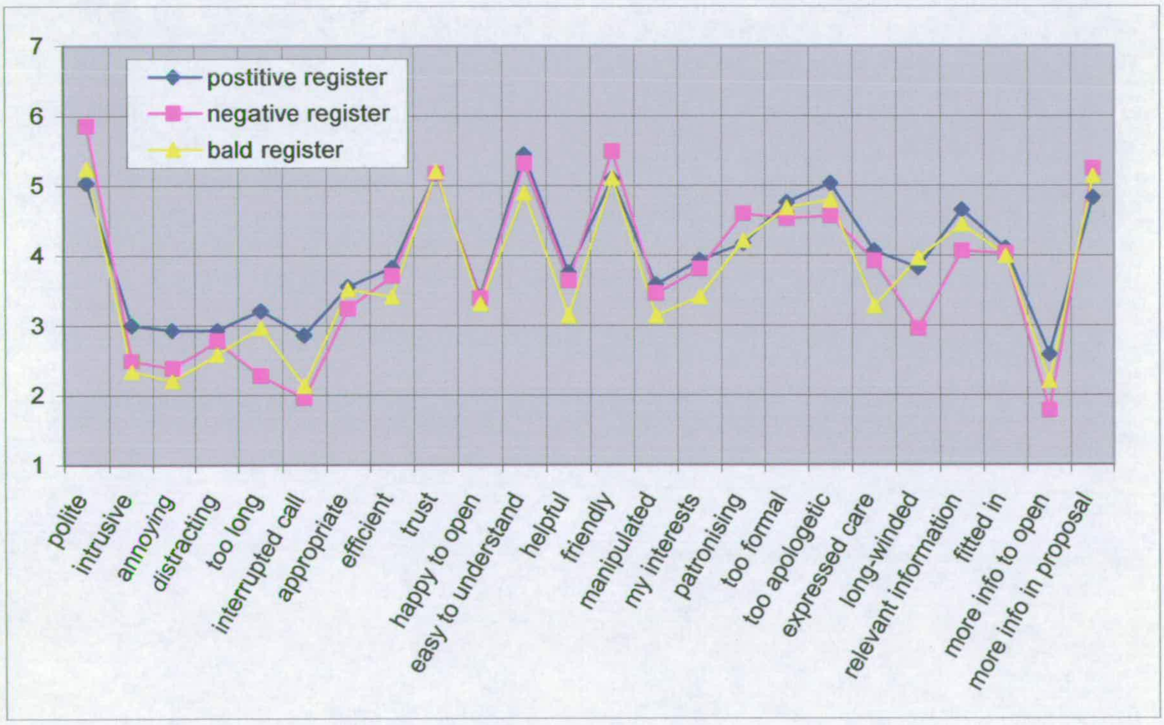


Chart 47. Mean score profiles (SIDQ0), split according to proposal politeness register, SID groups participants only (N = 86).

The Cronbach's Alpha for SIDQ0 was .91 (see section 3.10.5), which suggests a satisfactory level of consistency.

Participants completed a SID questionnaire after each phone call that included a SID delivery (three in total for each participant). The second set of analyses described in this section uses the pooled data from these questionnaires. A repeated-measures ANOVA was run on participant responses from SIDQ0-2 (after phone calls three, four and five) with age and gender as between-group variables (Table 61); the within-subject variable was SIDQ responses with three levels (Positive, Negative, Bald). Furthermore, the permutation order in which participants experienced the three SID offer registers was also included as a between-subjects factor (SID_ORD) with six levels. The main reason for including permutation order as a variable was participant preparedness for the proposal: upon hearing the first proposal delivery in the third call participants had not been primed about the pending proposal, nor had they been informed that they would be evaluating it. In subsequent phone calls, however, they were aware that a proposal would be played; they were prepared to listen to it and knew how they would be asked to evaluate it.

Repeated-measures analyses on pooled SIDQ0-2 mean scores revealed no effect of between-participant variables age, gender and SID order overall. There was, however, an overall significant main effect of the within-participant variable SID politeness register (SID_STRATEGY in Table 61). Tests of within-participant contrasts showed that this difference lay between the Positive face redressive style ($M = 3.83$) and the Bald offer strategy ($M = 4.22$), [$df = 1$, $F = 11.43$, $p = .001$]. The overall mean score for the Negative face redressive style was 4.06.

The analysis also revealed significant interaction between the participant SID register and age factors, as illustrated in Figure 22. To explore this interaction further, repeated-measures analyses were performed on the pooled SIDQ0-2 data for each individual age group separately (gender and SID order as between-participant variables); results showed statistical significance in the main effect of politeness strategy in the youngest age group [$df = 2$, $F = 13.197$, $p = .000$] and within-participant contrast revealed significant differences between the Positive face redress proposal and the Negative [*p* = .001] and Bald [*p* = .000] strategies (the Positive proposal receiving the lowest attitude ratings). No significant results were obtained in the analyses of responses from the mid- and older age groups. Univariate ANOVAs on the pooled data SIDQ0-2 (dependent variable) for each separate SID proposal condition (age, gender and SID order as independent variables) revealed no further significances.

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Between-participant effects					
AGE	2.304	2	1.152	1.634	.205
GENDER	.203	1	.203	.288	.594
SID_ORD	3.034	5	.607	.861	.514
AGE * GENDER	.512	2	.256	.363	.697
AGE * SID_ORD	6.101	10	.610	.865	.570
GENDER * SID_ORD	2.076	5	.415	.589	.708
AGE * GENDER * SID_ORD	5.676	10	.568	.805	.624
Error	35.244	50	.705		
Within-participant effects					
SID_REGISTER	3.889	2	1.944	4.629	**.012**
SID_ REGISTER * AGE	5.399	4	1.350	3.213	**.016**
SID_ REGISTER * GENDER	1.136	2	.568	1.352	.263
SID_ REGISTER * SID_ORD	10.616	10	1.062	2.528	**.009**
SID_ REGISTER * AGE * GENDER	1.044	4	.261	.621	.648
SID_ REGISTER * AGE * SID_ORD	12.146	20	.607	1.446	.119
SID_ REGISTER * GENDER * SID_ORD	2.043	10	.204	.487	.895
SID_ REGISTER * AGE * GENDER * SID_ORD	8.052	20	.403	.959	.517
Error (SID_ REGISTER)	42.001	100	.420		

Table 61. Univariate ANOVA on usability mean scores (pooled data from SIDQ0-2).

There was also significant interaction between factors politeness register and permutations of SID presentation order (Table 61), indicating the presence of an experiment order effect: one of the trade-offs associated with repeated-measures designs (see Section 3.8). Participant responses are illustrated in Figure 23. The order effect can be summarised as follows: attitudes to the Bald proposal are more positive when this is not the first exposure to a proposal. No additional statistical analyses were performed in order to explore this effect further, however, it is worth noting that participants' responses to the SID dialogue in the SIDQ measurement tool were influenced by exposure to contrasting politeness face redressive strategies in previous phone calls.

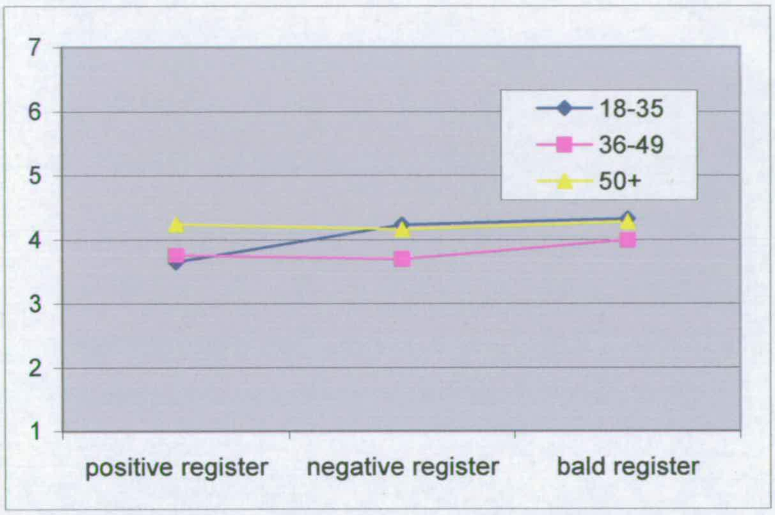


Figure 22. Overall mean usability scores (pooled data from SIDQ0-2) grouped according to SID register and age, all SID group participants (*N* = 86).

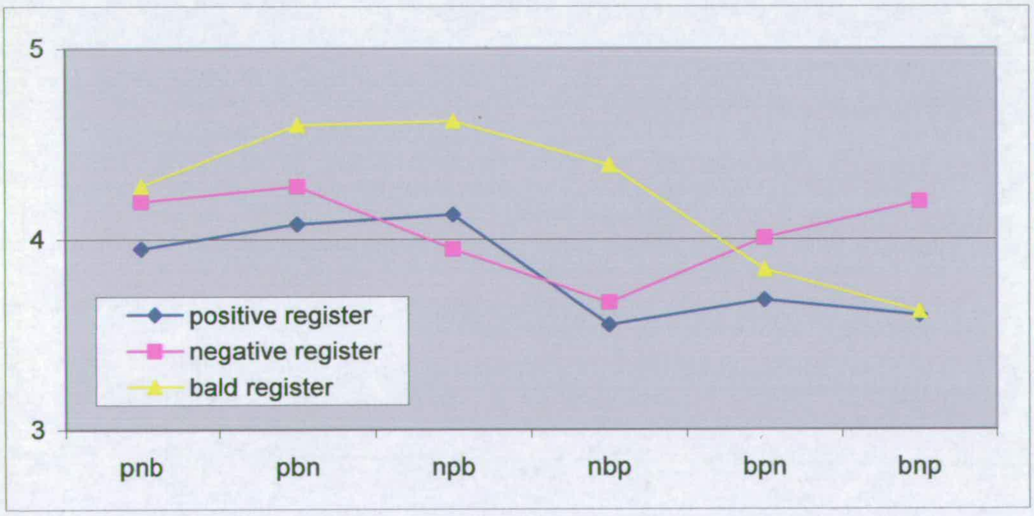


Figure 23. Overall mean usability scores (pooled data from SIDQ0-2) grouped according to SID register and SID order (p = Positive, n = Negative, b = Bald), all SID group participants (*N* = 86).

Repeated-measures ANOVAs (with between-group variables age, gender and SID order) were also performed on each of the individual pooled SIDQ0-2 statements; score profiles for the three face redressive styles are shown in Chart 48.

10 of the attributes in the pooled questionnaire responses showed significant differences between the three proposal registers (Chart 48). In terms of *politeness* [df = 2, F = 3.641, p = .030], participants rated the Negative register (within-subject contrasts) more polite compared to the Positive [df= 1, F = 6.737, p = .012] and Bald [df= 1, F = 4.316, p = .043] face redressive styles. There was significant difference between proposal registers for proposal *intrusiveness* [df= 2, F = 4.265, p = .017]; within-subject contrasts showing that the difference between the Positive and Bald registers was statistically significant [df= 1, F= 11.978, p = .001].

There were significant differences in terms of perceived *length* of the proposal [df= 2, F = 9.865, p = .000]; within-participant contrasts revealing that participants were significantly more positive to the length of the Bald register proposal compared to the Positive [df= 1, F = 18.000, p = .000] and Negative [df= 1, F = 8.636, p = .005] registers. For the *efficiency of using proposals to give users information* [df= 2, F = 3.172, p = .046] contrasts revealed that the difference between the Positive and Bald registers was significant [df= 1, F = 6.891, p = .011].

When it came to feeling *manipulated* by the proposal [df= 2, F = 5.683, p = .005], within-subject contrasts showed that the Positive register was rated significantly more manipulative compared to the Negative [df= 1, F = 5.500, p = .023] and Bald [df= 1, F = 13.151, p = .001] registers. The attribute *patronising* was also significant [df= 1.786, F = 7.477, p = .002]; the Bald proposal being rated significantly less patronising compared to the Positive [df= 1, F = 14.843, p = .000] and Negative [df = 1, F = 5.006, p = .030] face redressive styles.

Questionnaire item *the proposal was too formal* showed statistical significance [df= 2, F = 7.564, p = .001]; within-participant contrasts revealed that participants were more negative to the Negative face redressive register compared to the Bald [df= 1, F = 5.979, p = .018] and the Positive [df = 1, F = 13.687, p = .001] registers. Furthermore, in terms of the proposal being *too apologetic* [df= 1.683, F = 8.737, p = .001], the Negative register was rated the lowest (i.e. perceived to be too apologetic); the difference between the Positive and Negative registers was significant [df= 1, F= 11.201, p = .002].

The attribute regarding proposal *long-windedness* was significant [*df* = 2, *F* = 5.525, *p* = .005]; the Bald register proposal was rated more positively compared to both the Positive [*df* = 1, *F* = 7.017, *p* = .011] and the Negative [*df* = 1, *F* = 8.392, *p* = .006] face redressive styles. In terms of the proposal containing only *relevant information* [*df* = 2, *F* = 4.695, *p* = .011], participants were more positive to the Bald register compared to the Positive face redress [*df* = 1, *F* = 7.575, *p* = .008].

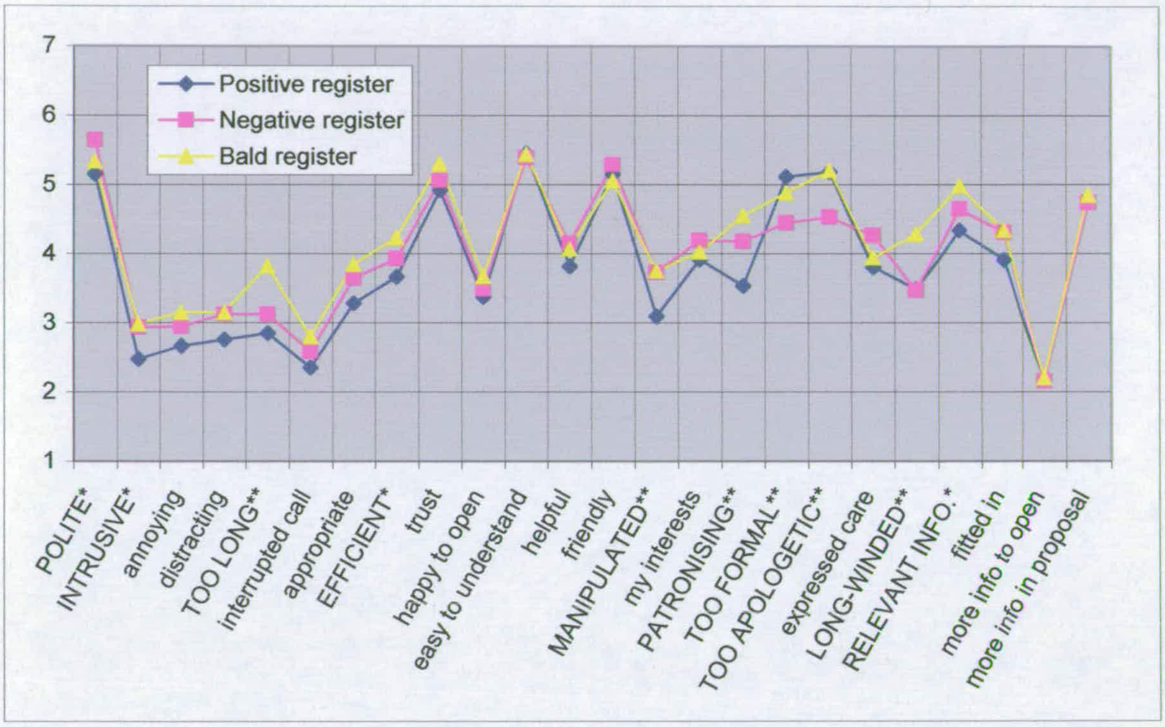


Chart 48. Mean score profiles (pooled data from SIDQ0-2), split according to proposal politeness register, SID groups participants only (*N* = 86), [*p<.05; **p<.01].

There were a number of interactions between the within-participant factor and the between-participant factors. These are summarised in Table 62.

In terms of between-participant effects, there were four items significant for effect of age: the perception of the proposal being *too formal* [*df* = 2, *F* = 8.130, *p* = .001], taking the individual's *interest into account* [*df* = 2, *F* = 4.533, *p* = .016], *expressing care for the individual's needs* [*df* = 2, *F* = 3.800, *p* = .029] and being *polite* [*df* = 2, *F* = 3.749, *p* = .031]. Post Hoc tests revealed that the youngest age group [*M* = 5.22] took a significantly more positive attitude to the proposals overall compared with the mid- [*M* = 4.63, *p* = .018] and oldest [*M* = 4.44, *p* = .000] age groups for perceived *formality*. For the item 'took my interest into account', the difference between the mid- [*M* = 3.55] and oldest [*M* = 4.53] age groups was significant [*p* = .030]; the mean for the youngest age group was 4.13. The item 'the proposal expressed care for my individual needs' showed statistical significance [*p* = .037] between the youngest [*M* = 3.81] and oldest [*M* = 4.30] age groups, while the difference between the mid- [*M* = 3.81] and oldest age groups approached statistical significance [*p* = .057]. In terms of the *politeness* of the proposal, the difference between the youngest [*M* = 5.22] and the oldest [*M* = 5.54] age groups was statistically significant [*p* = .037]; the mean for the mid-age group was 5.26.

Questionnaire item	Interaction	Statistics
The proposal was polite.	SID_REGISTER*AGE	*p*=.043
I found the proposal intrusive.	SID_REGISTER*SID_ORD	*p*=.004
The proposal was annoying.	SID_REGISTER*SID_ORD	*p*=.009
The proposal distracted me from what I was trying to do.	SID_REGISTER*SID_ORD	*p*=.018
The proposal was too long.	SID_REGISTER*SID_ORD	*p*=.017
The proposal interrupted the call too much.	SID_REGISTER*SID_ORD	*p*=.050
The proposal was appropriate for this service.	SID_REGISTER*AGE	*p*=.023
The proposal was an efficient way of giving information about the Online Saver account.	SID_REGISTER*SID_ORD	*p*=.045
The proposal information was helpful.	SID_REGISTER*SID_ORD	*p*=.000
The proposal was very long-winded.	SID_REGISTER*AGE	*p*=.029
The proposal took my interests into account.	SID_REGISTER*GENDER	*p*=.048
	SID_REGISTER*SID_ORD	*p*=.030
I found the proposal patronising.	SID_REGISTER*AGE	*df*=3.572 *p*=.011
The proposal contained only relevant information.	SID_REGISTER*SID_ORD	*p*=.047

Table 62. Summary of significant interactions for SIDQ0-2 (pooled data) usability attributes, for participants who had experienced a SID offer (*N* = 86).

Furthermore, two questionnaire items showed statistical significance in terms of SID order (SID_ORD): perceived proposal *formality* [*df* = 5, *F* = 3.312, *p* = .012] and finding the proposal *easy to understand* [*df* = 5, *F* = 3.552, *p* = .008]. There was a further two-way interaction (AGE*SID_ORD) for finding the proposal *distracting* [*df* = 5, *F* = 2.384, *p* = .021].

7.5.6 Perception of the politeness registers in the listening session (FRSQ)

In order to establish the absolute politeness in the proposal (i.e. attitudes toward the politeness strategies when removed from the context of the automated telephone dialogue), an additional session was included at the end of the experiment in which all participants (*N* = 111) listened to each proposal over computer speakers. Immediately after hearing a SID offer over the speakers, participants completed an FRS questionnaire (described in Section 7.4.5) where statements had been divided into two sets: perception of register employed and attitude to speaker characteristics. The data were then pooled according to politeness register employed (Positive, Negative and Bald) and repeated-measures ANOVAs were carried out on the individual questionnaire attributes.

The aim of the listening session was two-fold, to explore: (1) the participant's perception of the register and speaker characteristics employed in the contrasting proposals and (2) whether or not the contrasting face-redress strategies would produce effects consistent with Brown and Levinson's theories.

Before listening to the SID offers over the computer speakers, participants who had already experienced the SID offers in the automated service were told that they would hear each proposal again over computer speakers and that they would be asked to give their opinions about each message by completing some questionnaires. The control-group participants were instructed to imagine (before listening to the SID offers) that they had heard the following message in the automated service and also that after listening to the message they would be completing a questionnaire to give their opinions about it.

Firstly, analyses were performed on questionnaire responses from the control group only (*N* = 25); this most closely represents the 'absolute politeness' values of the SID offers as participants had not previously experienced the offers in the context of the automated service. Repeated-measures analyses were performed on the pooled response data (FRSQ0-2) with age, gender and presentation order (six levels) as between-subject variables. Some treatment cells had counts of 0; the GLM sums of squares were adjusted accordingly to Type IV.

Mean scores from the first set of questionnaire items (politeness register) are given in Figure 24. Three out of eight items showed statistically significant differences for main effects of SID offer (within-subject factor), these were: *informal/formal* [*df* = 2, *F* = 7.034, *p* = .010], *to the point/long-winded* [*df* = 2, *F* = 15.865, *p* = .000] and *apologetic/unapologetic* [*df* = 2, *F* = 8.443, *p* = .005]. Results from tests of within-participants contrasts for this set are presented in Table 63.

Four items were significant for between-participant effect of the SID order factor: *polite/impolite* [*df* = 5, *F* = 5.162, *p* = .035], *formal/informal* [*df* = 5, *F* = 6.841, *p* = .018], *sincere/insincere* [*df* = 5, *F* = 6.170, *p* = .023] and *apologetic/unapologetic* [*df* = 5, *F* = 12.229, *p* = .004]. Post Hoc tests revealed no further significant differences.

Three items were significant for between-participant effect of the gender factor: *formal/informal* [*df* = 1, *F* = 18.399, *p* = .005] with females finding the proposals more formal [*M* = 4.40] compared to males [*M* = 4.17]; *to the point/long-winded* [*df* = 1, *F* = 7.220, *p* = .036] with females finding the proposals overall more to-the-point [*M* = 2.93] compared to males [*M* = 3.17]; and *apologetic/unapologetic* [*df* = 1, *F* = 6.000, *p* = .050] with males [*M* = 3.73] finding the proposals overall somewhat more apologetic compared to females [*M* = 4.09]. There were no significant effects of age.

Similarly, repeated-measures ANOVAs were carried out on the second set of FRSQ statements which concerned perceived speaker characteristics. Mean scores are presented in Figure 25. Two out of six items showed statistical significance for main effects of SID offer (within-participant factor), these were: *tactful/tactless* [*df* = 2, *F* = 6.294, *p* = .014] and *professional/unprofessional* [*df* = 2, *F* = 15.050, *p* = .001]. Results from tests of within-participants contrasts for this set are presented in Table 64.

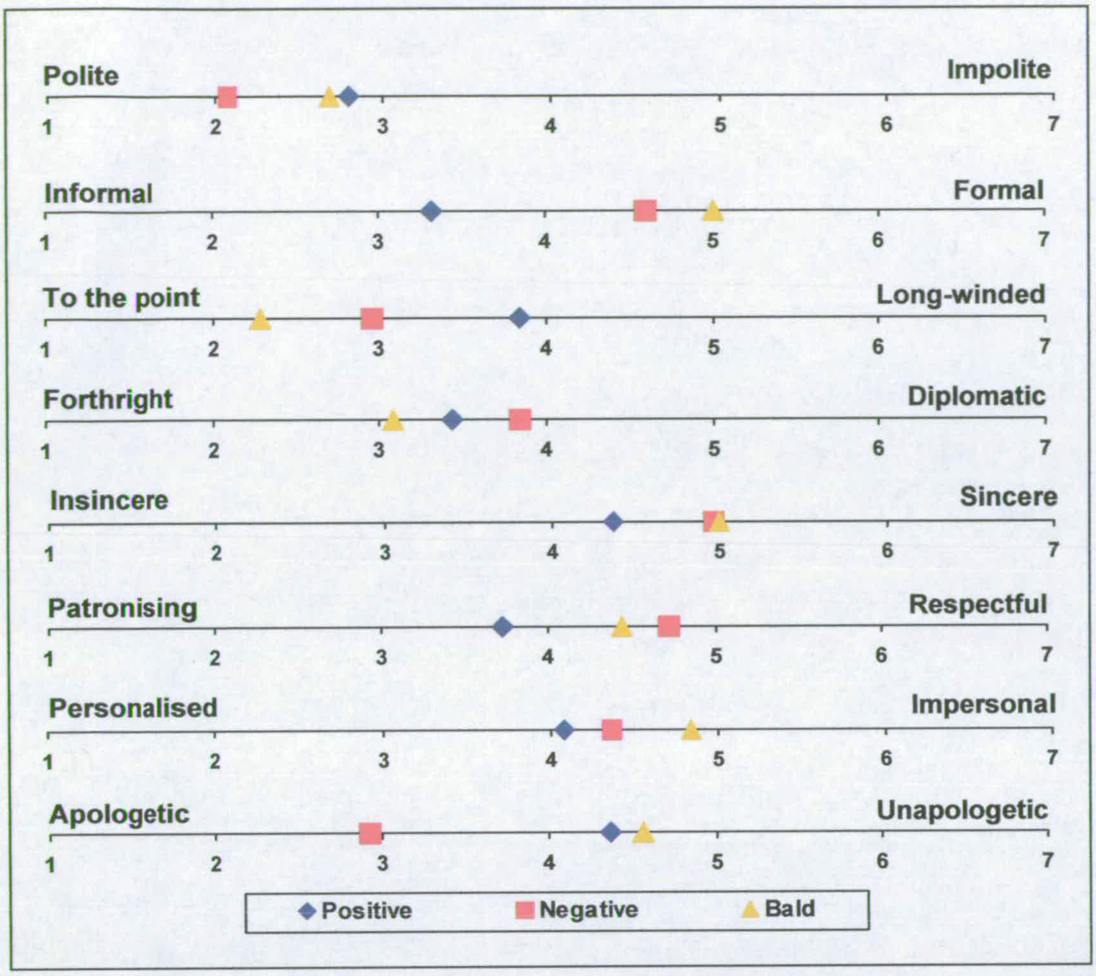


Figure 24. Control-group participant (N = 25) mean responses pooled according to politeness register (FRSQ0-2). These questionnaire items were introduced to participants with the phrase *"Thinking about the proposal I've just heard, it was…"* in the heading.

	Positive vs. Negative face redress *F=/p=*	Negative face redress vs. Bald strategy *F=/p=*	Positive face redress vs. Bald strategy *F=/p=*
Informal - Formal	5.348/.060	.681/.441	**24.174/.003****
To the point - Long-winded	**12.011/.013***	**9.054/.024***	**21.054/.004****
Apologetic - Unapologetic	**8.420/.009****	**9.338/.022***	2.625/.156

Table 63. Statistical results (contrasts) from the analysis of pooled mean responses (politeness register FRSQ0-2), for control-group participants (N = 25).

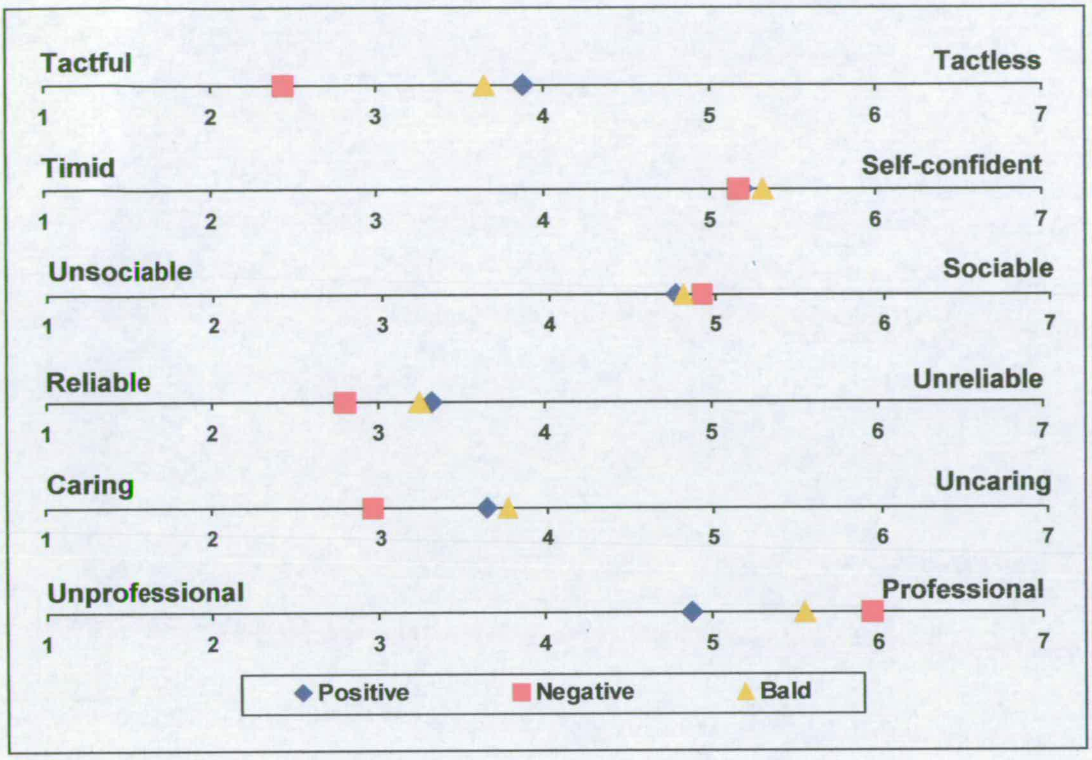


Figure 25. Control-group participant (N = 25) mean responses pooled according to politeness register (FRSQ0-2). These questionnaire items were introduced to participants with the phrase *"I associate the choice of wording in the proposal with someone who is…"* in the heading.

	Positive vs. Negative face redress *F=/p=*	Negative face redress vs. Bald strategy *F=/p=*	Positive face redress vs. Bald strategy *F=/p=*
Tactful - Tactless	**12.336/.013***	5.876/.052	.006/.941
Unprofessional - Professional	**21.429/.004****	**30.375/.001****	**6.482/.044***

Table 64. Statistical results (contrasts) from the analysis of pooled mean responses (speaker characteristics FRSQ0-2), for control-group participants (N = 25).

The second set of analyses (repeated-measures) was performed on pooled FRSQ0-2 questionnaire responses from all participants (N = 111) with age, gender and presentation order as between-participant variables. The benefit of using the whole participant sample was that larger samples generally produce more reliable statistical findings (see 3.10.2). A further between-participant variable (CONTROL_GRP) was included with two levels (yes/no) to account for the difference in experience between control group participants and the remaining participants (who had already heard the same proposals in the service). Again, some treatment cells had counts of 0 and consequently the GLM sums of squares were adjusted accordingly to Type IV.

Mean scores from the first set of questionnaire items (politeness register) are given in Figure 26. Six out of eight items showed statistically significant differences for main effects of SID offer (the within-participant factor), these were: *polite/impolite* [*df* = 2, *F* = 6.459, *p* = .002], *formal/informal* [*df* = 2, *F* = 12.405, *p* = .000], *to-the-point/long-winded* [*df* = 2, *F* = 19.693, *p* = .000], *forthright/diplomatic* [*df* = 2, *F* = 6.090, *p* = .003], *patronising/respectful* [*df* = 2, *F* = 7.170, *p* = .001] and *apologetic/unapologetic* [*df* = 1.814, *F* = 33.358, *p* = .000]. Results from tests of within-participants contrasts for this set are presented in Table 65.

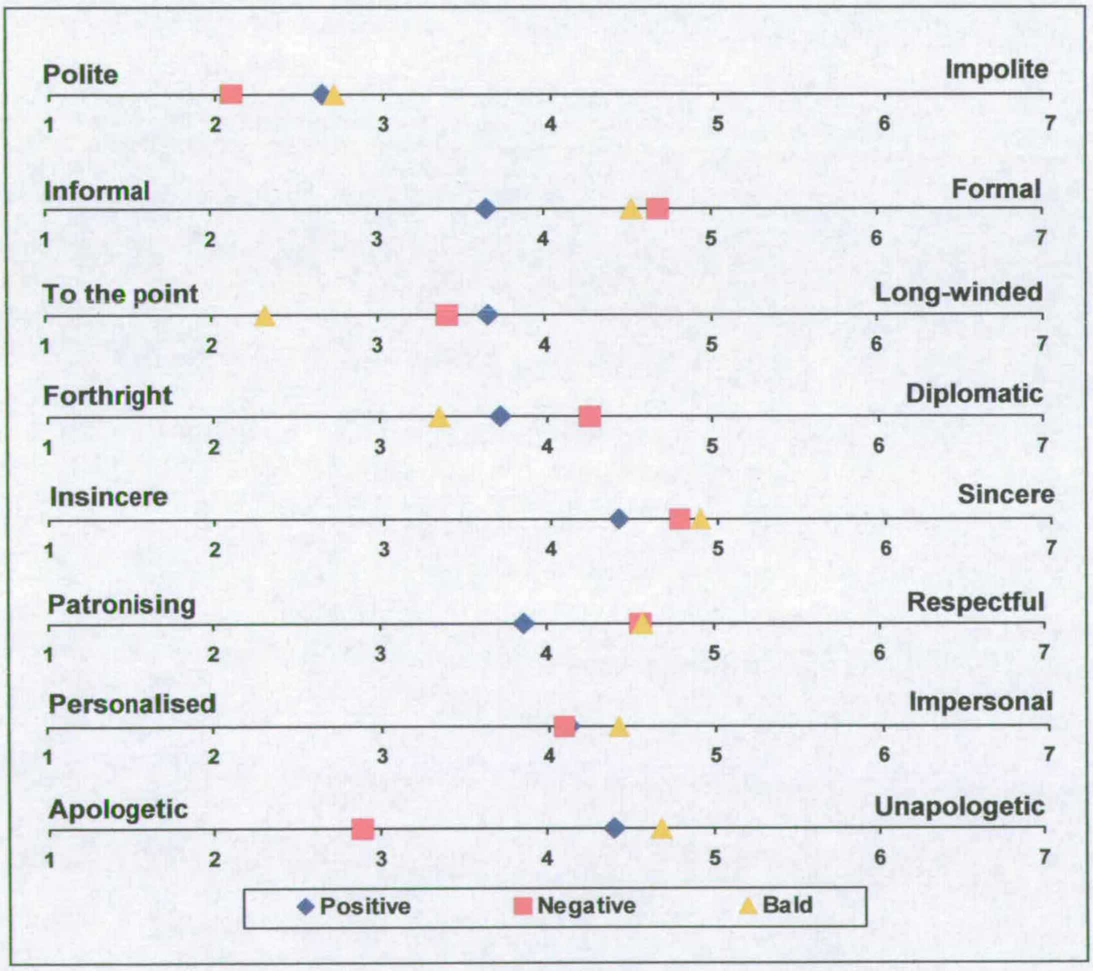


Figure 26. Mean responses (all *N* = 111 participants) pooled according to politeness register (FRSQ0-2). These questionnaire items were introduced to participants with the phrase "Thinking about the proposal I've just heard, it was..." in the heading.

	Positive vs. Negative face redress F=/p=	Negative face redress vs. Bald strategy F=/p=	Positive face redress vs. Bald strategy F=/p=
Polite - Impolite	**9.928/.003****	**11.662/.001****	.370/.545
Informal - Formal	**17.397/.000****	.018/.895	**23.821/.000****
To the point - Long-winded	2.626/.110	**21.011/.000****	**35.856/.000****
Forthright - Diplomatic	1.122/.294	**12.230/.001****	**5.612/.021***
Patronising - Respectful	**15.140/.000****	.513/.477	**6.958/.011***
Apologetic - Unapologetic	**35.870/.000****	**48.016/.000****	3.404/.070

Table 65. Statistical results (contrasts) from the analysis of pooled mean responses (politeness register FRSQ0-2), whole participant cohort (*N* = 111).

Main effects for between-participant factor age were obtained for item *polite/impolite* [*df* = 2, *F* = 4.323, *p* = .018], Post Hoc tests revealing that the oldest age group [*M* = 1.90] thought the proposals were significantly more polite overall compared to the youngest [*M* = 2.69, *p* = .001] and the mid-age groups [*M* = 2.61, *p* = .010]. In terms of the item *to-the-point/long-winded* [*df* = 2, *F* = 4.264, *p* = .018], the oldest age group [*M* = 2.41] found the proposals overall more to-the-point compared the youngest [*M* = 3.44, *p* = .001] and the mid-age groups [*M* = 3.25, *p* = .017]. The item *sincere/insincere* was also significant for age [*df* = 2, *F* = 7.591, *p* = .001]; again the difference lay with the oldest age group [*M* = 5.23] rating the proposals overall more sincere compared to the youngest [*M* = 4.51, *p* = .007] and the mid-age groups [*M* = 4.52, *p* = .014]. The item *patronising/respectful* [*df* = 2, *F* = 4.561, *p* = .014] was also significant [*M* 18-35 = 4.31; *M* 36-49 = 4.10; *M* 50+ = 4.63] however Post Hoc tests revealed no further differences.

One questionnaire item – *formal/informal* – showed statistical significance for the gender factor [*df* = 1, *F* = 7.841, *p* = .007] with female [*M* = 4.42] participants finding the proposals overall more formal than males [*M* = 4.10]. The item *forthright/diplomatic* was significant for the proposal group factor [*df* = 1, *F* = 6.391, *p* = .014] participants who were in the control-group [*M* = 3.45] finding the proposals overall more forthright compared to the remaining participants [*M* = 3.87]. There was also one significant item *patronising/respectful* for factor SID presentation order [*df* = 5, *F* = 3.297, *p* = .011]; Post Hoc tests revealed no further significant differences.

Mean scores for the second set of FRSQ statements concerning perceived speaker characteristics are presented in Figure 27. Repeated-measures ANOVAs revealed that four out of six items were significantly different for main effect of SID politeness (within-participant factor), these were *tactful/tactless* [*df* = 2, *F* = 16.119, *p* = .000], *timid/self-confident* [*df* = 2, *F* = 7.246, *p* = .001], *caring/uncaring* [*df* = 2, *F* = 7.926, *p* = .001] and *professional/unprofessional* [*df* = 2, *F* = 10.658, *p* = .000]. Results from tests of within-participants contrasts for this set are presented in Table 66.

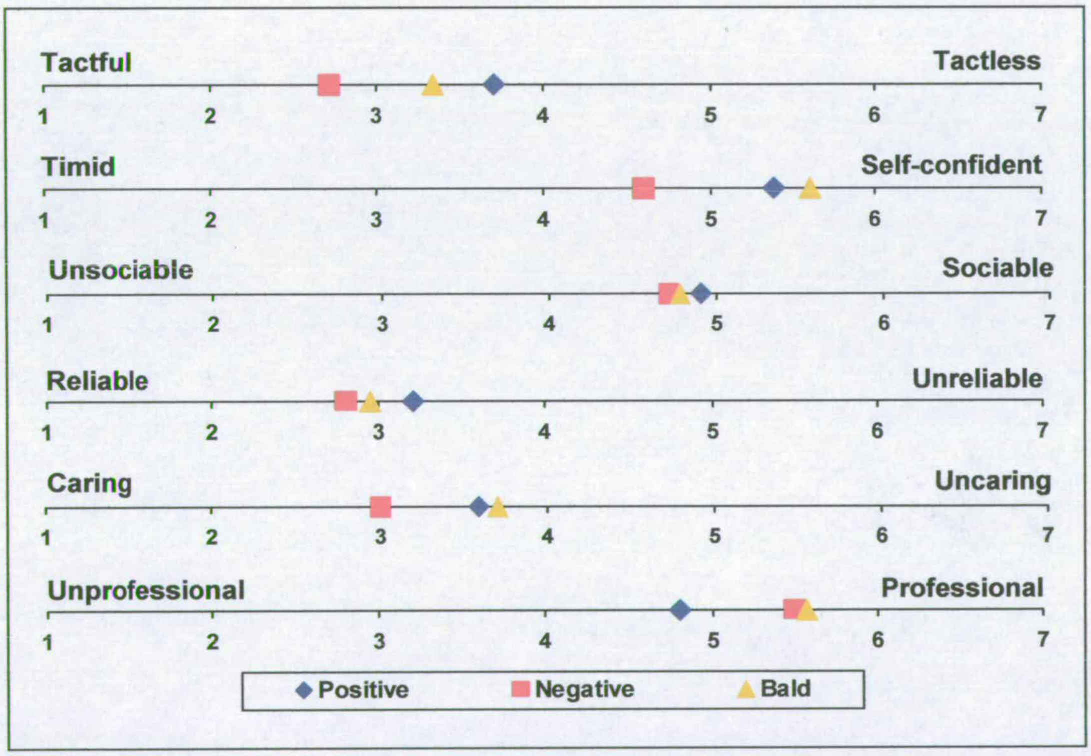


Figure 27. Mean responses (all *N* = 111 participants) pooled according to politeness register (FRSQ0-2). These questionnaire items were introduced to participants with the phrase *"I associate the choice of wording in the proposal with someone who is…"* in the heading.

Three items were significant for main effect of age for the item *tactful/tactless* [*df* = 2, *F* = 7.631, *p* = .001], Post Hoc tests revealed that the oldest age group [*M* = 2.54] considered the speaker overall more tactful compared to the youngest [*M* = 3.40, *p* = .001] and mid-age group [*M* = 3.63, *p* = .000]. The item *reliable/unreliable* [*M* 18-35 = 3.00; *M* 36-49 = 3.17; *M* 50+ = 2.73] was also significant [*df* = 2, *F* = 3.980, *p* = .024] however Post Hoc tests revealed no further differences. In terms of *caring/uncaring* [*df* = 2, *F* = 9.370, *p* = .000], Post Hoc tests revealed significant difference [*p* = .003] between the oldest [*M* = 3.02] and mid-age groups [*M* = 3.73]; the mean for the youngest age group was 3.46.

	Positive vs. Negative face redress *F=/p=*	Negative face redress vs. Bald strategy *F=/p=*	Positive face redress vs. Bald strategy *F=/p=*
Tactful - Tactless	**36.065/.000****	**18.880/.000****	.687/.410
Timid - Self-confident	**5.621/.021***	**12.681/.001****	1.831/.181
Caring - Uncaring	**9.201/.004****	**15.668/.000****	1.173/.282
Unprofessional - Professional	**22.953/.000****	.232/.631	**10.574/.002****

Table 66. Statistical results (contrasts) from the analysis of pooled mean responses (speaker characteristics FRSQ0-2), for all participants (*N* = 111).

There was no indication in the analyses that previous experience of the proposals in the automated service (i.e. the CONTROL_GRP variable) had any impact on participant attitudes to the proposal in the listening session. In line with the findings from the analysis of SIDQ, there was also some evidence that presentation order had an impact on the participants' responses; there was an interaction between the within-participant variable SID politeness register and the between-participant variable SID order for items *to-the-point/long-winded* [*p* = .034], *patronising/respectful* [*p* = .036], *personalised/impersonal* [*p* = .011], *timid/self-confident* [*p* = .043] and *reliable/unreliable* [*p* = .003].

7.5.7 De-briefing interview feedback

All participants took part in a structured interview after their experience with the automated service in which they were asked to compare the three SID proposals. To aid memory and to facilitate reference to the contrasting politeness strategies, three sheets containing the wording of each proposal were given to participants. Firstly participants were asked: which of the proposals do you prefer? Their responses are presented in Table 67.

Group	Positive	Negative	Bald	None	Unsure	Other	*N* =
Control Group	4 (16.0%)	11 (44.0%)	9 (36.0%)	-	-	1 (4.0%)	25
SID Groups	8 (9.3%)	21 (24.4%)	51 (59.3%)	5 (5.8%)	-	1 (1.2%)	86
All	12 (10.8%)	32 (28.8%)	60 (54.1%)	5 (4.5%)	-	2 (1.8%)	111

Table 67. Participant responses to "which of the proposals do you prefer?". In the 'Other' category: one participant chose Positive and Negative; another chose Negative and Bald.

Participants were then asked to express the reasons for their choice. There was no consensus about their preferred choice among participants who said they preferred the Positive face redress. Examples of comments were: "more positive", "more caring", "more polite" and "not so apologetic". The majority of participants who chose the Negative face redress style said the preferred it because it was "polite" or "apologetic"; some participants said they liked expressions such as "sorry to interrupt", "bank's policy" and "happy to set up". Some participants who chose the Bald style of proposal expressed dislike for the Positive face redress, others found they preferred the Bald style as it was "more short" and "more to-the-point" than the other two; further comments from this group were that the Bald style was "less patronising", "less intrusive", "less formal", "more honest" and "more professional".

Responses very much similar to those presented in Table 67 were also obtained when participants were asked "which of the proposals do you think is most suitable for an automated telephone banking service?". In the control-group there was a slight preference for the Negative face redress while participants who had experienced the proposals in the automated service chose the Bald style.

Participants were also asked which of the proposals they found to be most polite (Table 68). Examples of comments from participants who thought the Positive face redress was the most polite way to address the caller were: "more familiar", "more natural" and "not as blunt nor as apologetic as the other two". Most of the comments from participants who chose the Negative face redress as the most polite proposal picked up on the apology in the opening statement. The participants who chose the Bald style did so because it was shorter, more to-the-point and contained less unnecessary wording; they also found it less patronising and less apologetic.

Group	Positive	Negative	Bald	None	Unsure	Other	*N* =
Control Group	3 (12.0%%)	18 (72.0%)	3 (12.0%)	-	1 (4.0%)	-	25
SID Groups	11 (12.8%)	55 (64.0%)	17 (19.8%)	1 (1.2%)	-	2 (2.3%)	86
All	14 (12.6%)	73 (65.8%)	20 (18.0%)	1 (0.9%)	1 (0.9%)	2 (1.8%)	111

Table 68. Participant responses to "which of the proposals is the most polite way to address the caller?". In the 'Other' category: one participant chose Positive and Negative; another chose Positive, Negative and Bald.

Participants who thought the Positive face redress was the least polite (Table 69) mainly commented on the opening statement – "know you won't mind" – which they perceived as presumptuous. The main concern among participants who chose the Negative face redress as the least polite was that they thought the statement "bank's policy" indicated less concern about the actual customer's finances. Participants who chose the Bald style did so because they found it to be "abrupt" and "too forthright"; unsurprisingly it was the opening statement – "I'm interrupting" – in the proposal which elicited a strong reaction in this case.

Group	Positive	Negative	Bald	None	Unsure	Other	*N* =
Control Group	11 (44.0%%)	-	9 (36.0%)	3 (12.0)	1 (4.0%)	1 (4.0%)	25
SID Groups	42 (48.8%)	6 (7.0%)	30 (34.9%)	5 (5.8%)	2 (2.3%)	1 (1.2%)	86
All	53 (47.7%)	6 (5.4%)	39 (35.1%)	8 (7.2%)	2 (1.8%)	2 (1.8%)	111

Table 69. Participant responses to "which of the proposals is the *least* polite way to address the caller?". In the 'Other' category: the two participants chose both Positive and Bald.

Furthermore, participants were asked if they could perceive any benefits in having Online Saver proposals in this kind of service. The majority (*N* = 79, 71.2%) of participants responded "yes", some of them adding that there are some benefits in receiving information about new services and in hearing information about better interest offers and that they thought the automated service was a better channel for disseminating information compared to receiving information through the post (which many said they never read). The ability to reach customers who infrequently visit the branch was also mentioned. Participants who answered "no" commented that they did not think that these kinds of offers should be made through the automated telephone banking service.

When asked whether or not the proposals could be improved in any way, 71.2% (*N* = 79) thought that they could. Only a few comments were made on improving individual proposal styles, but the majority of comments were regarding changes that applied to all the proposal styles. For example, a significant number of participants said they would rather the proposal was at the end of their banking enquiry. There were also comments regarding the options that the proposal presents: rather than asking "would you like to set up an Online Saver account now?" participants wanted the option to skip the account information or to have an option to hear more information about the account before setting it up. Other participants would like the option for an Online Saver information pack to be sent to them through the post or the ability to be transferred to a human agent. Among other suggestions for improvements were shorter and more customised information.

7.6 Conclusions

The research in this chapter centred around three main themes:

1) presence/absence of SID with explicit interruptions,

2) attitudes to contrasting face redressive styles in the proposals and

3) evaluation (manipulation check) of SID politeness strategies when heard in isolation.

The key issue examined in this research is how politeness strategies (considered an important factor in the choice of vocabulary in human-human dialogue interruptions) may be employed to influence the impact of such system-initiated digressive proposals on user attitudes.

7.6.1 The impact of presence of SID

The prediction for this experiment was that the presence of a SID delivery with explicit interruptions in the opening phrase would have negative impact on user attitudes toward the usability of the service. Three contrasting politeness strategies (Positive, Negative and Bald), derived from established face-redress theories in human-human communication, were employed in order to mitigate the adverse effects of these dialogue intrusions.

The analysis of differential scores (UQ1-UQ0) for the whole participant group revealed that the change in attitude overall (Intercept) was significant and further that there was a main effect for presence of SID: participants who experienced a SID offer took a significantly more negative attitude to service usability compared to the control group. This verifies the experiment prediction that the SIDs would have a negative impact on service usability.

Furthermore, the majority of individual statements (13 out of 20) were significant in terms of presence/absence of a proposal; again, the change in attitude for participants who experienced a SID being significantly more negative compared to the control group. Items that remained unaffected were *speed* and *reliability* of the service, the *liking* and *clarity* of the voice, and *preference for speaking to a human*; perhaps more surprisingly there were no difference between the two groups in terms of perceived *politeness* and *friendliness* of the service.

To further explore the change in attitude and to compare the three contrasting SID politeness strategies, analyses were performed on differential scores (UQ1-UQ0) from the participant subgroup that had experienced a SID offer. Confirming the results in the preceding paragraphs, the change in attitudes overall (Intercept) was significant and the negative impact in attitude scores was significant for 16 out of 20 items (non-significant items were *liking* and *clarity* of the voice, *friendliness* of service and *concentration*).

In sum, in accordance with experiment prediction 1, the presence of a savings offer containing an explicit interruption had a significant negative impact on user attitudes to the service.

7.6.2 The impact of contrasting SID politeness register

No specific predictions had been made regarding the impact of contrasting politeness registers on user attitudes to the usability of the automated service. Based on the analysis of the differential scores (UQ1-UQ0), there was little indication that employing contrasting politeness registers had any effect on user attitudes: there were no significant differences overall for the SID register factor and only two individual questionnaire items (*ease of use* and *reliable*) were moderately significant ($p < .05$). This indicates that the negative impact of the proposals on perceived service usability was comparable and consistent across all three strategy groups. Interestingly, the results show that the types of apology and politeness used in the Negative face-redress strategy (which are typically associated with politeness etiquette) were not effective. The use of "I'm very sorry to interrupt..." in the Negative face-redress was no better received than the phrase "I'm interrupting..." in the Bald strategy.

The current experiment included a set of 24 questionnaire statements (SIDQ) aimed at capturing participants' attitudes towards the SID dialogue directly. It was predicted that varying the politeness strategy of the SID offers would have some impact in terms of perceived social appropriateness (the Negative strategy being more highly rated in terms of politeness) and that the variation in duration between the SID messages would result in the Bald strategy being more positively received.

Two analyses were performed on the SIDQ response data. Firstly, responses to first exposure to a SID offer (SIDQ0) were examined. Results showed no significant differences overall for the SID strategy factor, nor were there any significant differences in the analysis of individual questionnaire items. It is worth noting that approximately half of the mean scores for individual statements fell below the neutral 4-point mark. In particular, attributes relating to the intrusiveness and disruptiveness of the proposal (i.e. *annoying, too long, intrusive, distracting* and *interrupted the call*) generated negative responses with mean scores around or below 3.0. The conclusion is then that participants' reactions to the three politeness strategies, albeit contrasting registers employed, were negative and consistent after first exposure.

The analysis of the pooled response data (after participants had experienced all three proposals) revealed that there were significant differences overall between the Bald and the Positive face-redress strategies (in favour of the Bald strategy). In a real world scenario, users would not experience the full range of proposal strategies in this manner, however, the results from the pooled analysis provide some qualitative guidance on the design issues involved in attempts to add such digressions to automated telephone dialogues by eliciting participants' preferences for the wordings of such proposals. As predicted, participants rated the Negative face redressive style as being the most *polite* of the three; however, there was also indication that politeness had been taken 'to the extreme' in this case as participants also found the Negative style to be *too formal* and *too apologetic*. The Bald strategy received significantly more positive response in terms of being *shorter* and *less long-winded* compared to the two other face redressive styles. Furthermore, the Bald strategy was perceived to be *least patronising* of the three proposals and, when compared with the Positive politeness strategy, was judged to contain more consistently *relevant information*, was *more efficient* and *less intrusive*.

To conclude, the analysis of participant responses from first exposure to the three proposal styles does not provide any supporting arguments for selecting one politeness strategy over the other. The consistently negative reaction to the SID deliveries firmly supports the avoidance of starting a proposal with an explicit interruption. An interesting extension to the current experiment would be to explore participants' reactions to two versions of the Bald proposal strategy: one version which retains the wording used here, and one modified version which replaces the initial interruption phrase with "you might like to know that...". This would provide an indication of how sensitive users are to the opening phrase used in SIDs.

The analysis of the pooled SIDQ responses has been useful in terms of highlighting user preferences for the wording of the proposals which in turn may be used by the dialogue engineer in prioritising design criteria. The findings from the pooled data are supported by participant comments from the de-briefing interview: the Bald proposal preferred because it was shorter and more to-the-point. However, having actually heard the proposal in the automated service also seems to have an impact on user responses: the most preferred proposal strategy in the control group was the Negative (44.0%) while among remaining participants the Bald strategy was preferred (59.3%). This suggests that there are qualities in the Negative face redress that appeal to users, but also stresses that the perception of proposed SID designs may be perceived and rated differently whether experienced 'hands on' in the context of the automated service or by 'face value' in a listening session.

Furthermore, the de-briefing interview also served to show the interpretation of the concept of 'politeness'. The Negative face redressive style was, as expected, rated to be the most polite of the three proposals (65.8%) – mainly due to the apology in the opening statement. In terms of the *least* polite proposal, 47.7% of participants chose the Positive face redressive style and 35.1% the Bald strategy. Again, it was the opening statement which participants reacted to most strongly: the Bald "I'm interrupting" was perceived as abrupt and the "know you won't mind" as presumptuous further indicating the importance of getting the initial approach right in order to construct an acceptable SID offer.

In sum, in accordance with experiment prediction 2, the pooled data analysis showed some evidence that the Negative face redressive strategy was perceived to be the most polite of the three; the Bald strategy was rated to be shorter, more to-the-point and less intrusive. These differences, however, were not prominent based on first exposure to a proposal where there were no significant differences in the results.

7.6.3 The effect of age on attitudes to the SID offer

The analysis of change in attitude to service usability (UQ1-UQ0) showed an interaction overall (moderate effect, p = .047) between age and SID register: the youngest and mid-age group responded more consistently to the three proposal strategies compared to the oldest age group. Within the oldest age group the Positive face redressive proposal had the least negative impact on usability ratings overall followed by the Negative and Bald (most negative impact) proposal strategies; the difference between the Positive and Bald proposals was statistically significant. There was, however, no evidence that the oldest age group were more positive in their attitudes to the Positive face redress when they were asked to evaluate the three proposals (SIDQ).

Further evidence of an interaction between age and SID register appeared in the analysis of the pooled SIDQ (SIDQ0-2) responses. When analysing age groups separately, the difference in attitude to the three proposal registers appeared to be mainly within the youngest participants who took a significantly lower attitude to the Positive face redressive style overall compared to the other two strategies.

These findings are interesting and suggest that different age groups may react differently to varying politeness registers, both in terms of their relative impact on service usability and how they are rated. The evidence here, however, is inconclusive and further studies are needed in order to explore the link between age differences in linguistic theory and voice interfaces in human-computer interaction.

7.6.4 Task completion

Somewhat unexpectedly, seven out of the 29 subjects in the Bald proposal group failed to carry out their balance task after hearing the proposal. Instead of responding "yes" to the system prompt "would you like another service?" these participants said "no" and their phone calls were subsequently ended. Unfortunately more detailed information as to why participants said (wrongly) "no" had not been captured during the experiment session and therefore no definite conclusions regarding their reasoning could be made. A plausible explanation is that the way the Bald strategy was worded threw participants into confusion about what "another service" in the ensuing question actually referred to; alternatively, the blunt and concise register employed in the Bald proposal could have been particularly distracting to participants causing them to lose focus on their task at hand.

Irrespective of the cause, this sort of side effect can potentially hamper the use of the automated service and should be avoided at all costs. For SID offers to be successfully deployed in the existing automated service, callers must be able to carry out their main objective with the call without too much disruption; having to phone back and carry out the identification and verification dialogue again is a serious inconvenience. Therefore, the dialogue engineer must carefully consider the impact of a proposal, not only on attitudes and preference, but also on the dialogue experience as a whole. The cause of any undesirable side effects, such as reduced task completion, must be identified and resolved.

7.6.5 Attitudes to SID registers in isolation (listening session)

The three proposals were devised according to Brown and Levinson's politeness theory and the main purpose of the listening session was to provide a 'manipulation check' of the politeness registers employed. Broadly, Negative face redress is what we normally refer to as being 'polite'; the Positive face redress assumes a more intimate relationship between the interlocutors; and the Bald strategy does not adhere to any mitigating registers.

The data were analysed both for the control group only and then for all participants. The pooled analysis included a two-level factor for previous exposure (i.e. control group vs. not control group); as there was no evidence of previous exposure having any effect on participant ratings in the analysis the discussion here will focus on the results from the pooled data analysis.

In brief, significant differences between the proposal prompt registers employed were revealed. Out of the eight items aimed at exploring the choice of wording four more or less followed the predicted rating (in terms of significance for the questionnaire items *polite/impolite, formal/informal, long-winded/to-the-point, apologetic/unapologetic*). Not surprisingly, the Bald strategy that was perceived as significantly more *forthright* compared to either of the face redressive strategies and the Positive politeness register was rated as significantly more *patronising* than the other two SID offers. More surprising was the fact that participant – along the scale *polite/impolite* – rated all three politeness as being 'polite'. Similarly, against the prediction, the Bald strategy was rated equally *respectful* to the Negative face redressive style.

In terms of speaker characteristics, the Negative politeness register was rated (as expected) to be most *tactful* and *timid* of the three. The Negative politeness strategy was rated to be significantly more *caring* than the other two offer strategies and the Positive politeness strategy as the most *unprofessional* of the three. No significant differences were found for attributes *sociable/unsociable* or *reliable/unreliable.*

The results confirm that contrasting proposal registers may convey information beyond the content level, such as manners and speaker characteristics in system prompts. However, the listener's resulting interpretation of the level of 'politeness', 'respectfulness' and 'professionalism' in the message will depend on the context in which it is being used, social distance between interlocutors, personal expectations and preferences. Subsequently, the goal of devising a 'universal' taxonomy of politeness with associated production rules appears unattainable. However, Brown and Levinson's theory is well established within the field of human-human communication and there is opportunity for further studies to explore and define a similar taxonomy over 'polite' behaviour in spoken human-computer interaction.

In sum, in accordance with experiment prediction 3, participants perceived differences in the three contrasting politeness registers used and the ratings given by participants were, by and large, in line with Brown and Levinson's theories.

Chapter 8

We shall never cease from exploration
And the end of all our exploring
Will be to arrive where we started
And know the place for the first time.

- T. S. Eliot (1888 - 1965) -

General discussion and conclusions

8.1 Introduction

Today's speech-enabled automated telephone services typically rely on menu listings as a means of letting callers know about the range of options from which they may choose. An issue pertaining to the development and maintenance of such automated services is where to suitably introduce new or less frequently requested options, an area which has not yet been fully addressed in the current literature. The solution proposed in the current study involves the use of SIDs (system-initiated digressions) to inform callers about the availability of a new service option.

In a series of four experiments, SIDs featuring contrasting strategies, locations and registers were deployed in the dialogue of an existing mass-market automated telephone banking service. Participants were invited to try the automated service and their attitudes to the service (both before and after exposure to the digression) and to the SID dialogue itself were assessed by means of questionnaires. In addition, issues regarding task completion and participants' abilities to successfully locate and select the new (hidden) menu option were explored.

Voice user interfaces are a relatively new field within human-computer interaction research. The purpose of the current research has been to further the knowledge in this field by identifying dialogue engineering strategies for SIDs and evaluating the usability (effectiveness, efficiency and satisfaction) of such novel dialogue behaviour. The target application in the current research was an automated telephone banking application, however, the findings obtained in the experiments extend beyond the financial sector and are relevant to the design of a wide range of menu-driven self-service automated telephone services.

The findings from the current research have laid bare some interesting facts in terms of the usability and users' perceptions of SID dialogues in menu-driven automated telephone services, but have also generated further questions regarding users' abilities to operate such applications, the use of politeness and whether or not alternative SID delivery methods could be employed. The purpose of this section is to summarise and highlight these issues, and to propose research directions for further study.

8.2 Main findings and future research

The reasoning behind the SIDs was to introduce information about new service options within the automated service, without making any potentially costly changes to the core system dialogue (i.e. the existing flows, prompts and voice recordings). The purpose of the digression was to function as an independent component that could be 'bolted on' to an existing service, providing that recognition grammars are updated to process any caller speech input induced by the information in the digression. Three main dialogue engineering issues pertaining to SID dialogues were explored in the current research: strategy, location and register.

8.2.1 SID strategy

System prompts in automated telephone services mainly take two forms: informational messages (where the system retains the turn-taking initiative) or requests for a response from the caller (where the turn is handed over to the caller). Subsequently, two SID dialogue strategies were identified: a Signpost offer (informational message) and a Follow-on offer (informational message followed by a yes/no question for pursuing the offer).

These strategies were evaluated in two experiments (Experiment 1 and Experiment 3) where overdraft offers were deployed immediately after the participant had completed a balance request. Reassuringly, it was found that the digressions (irrespective of strategy employed) did not have a negative impact on user attitudes to the usability of the automated service. This leads to the conclusion that system-initiated digressions can successfully be deployed in mass-market automated services.

In terms of a 'winning' strategy, there were no strong findings which suggested user preference for either the Signpost or Follow-on digression. The Signpost digression was perceived to some extent more positively in terms of it being shorter and less intrusive; furthermore, in the de-briefing interview, participants who experienced the Signpost strategy were more positive to experiencing a SID offer in real use of the automated service compared to the Follow-on participant group. The results suggest a slight preference for adoption of the Signpost strategy. However, the extra turn-taking stage involved in the Follow-on dialogue does not have a significantly negative impact on user attitudes. The conclusion is therefore that, where it is appropriate to prompt the caller for a response in connection with a digression, the Follow-on strategy can also be adopted.

The purpose and content of the SID offer may suggest that alternative variations of the strategies defined above should be employed. For example, the service provider may want to ascertain that the callers understand what a product (such as "overdraft") actually involves initially by using yes/no dialogues such as "are you interested to hear more about...?". Depending on the amount of information associated with the product or service (such as terms and conditions) the service may offer the caller to "hear" it over the phone, "receive an information pack through the post" or "speak to a human". In some cases it may be necessary to include specific details where names of novel products and services are defined for the first time (such as for the "Online Saver" account introduced in the current research). The primary aim in such cases should be to include only the most pertinent information.

8.2.2 SID location

The dialogue in an automated telephone banking service had three main parts: an initial caller greeting, an identification and verification process and, finally, the main menu listing which gives the caller access to individual service options. The SID could therefore suitably be located immediately after the initial greeting ('Welcome'), after successful verification of the caller's identity ('ID&V') or following a completed transaction ('Transaction').

These three SID locations were evaluated in Experiment 2 by deploying Signpost-strategy overdraft offers. The results re-confirmed that the presence of a SID did not have a significant impact on user attitudes towards service usability. Furthermore, there were no significant differences between the three offer locations (evaluated both in terms of impact on service usability and perceptions of the SID offer itself) suggesting that any of these locations would be, more or less, suitable for a SID delivery. In the de-briefing interview, participants who experienced a proposal were somewhat more positive to the idea of receiving an offer in real-life use of the service, however, this is not enough evidence to suggest that the Welcome SID location would be the most suitable.

In fact, trade-offs associated with the Welcome SID location are likely to outweigh any positive attributes. One consequence of approaching callers prior to identification is that they would hear the SID offer every time they contacted the service; although not pointed out by participants in the experiment, this is likely to be perceived as annoying, particularly to frequent users. Furthermore, the Welcome location is unsuitable for product or services where only a subset of callers is eligible – as shown in the current research, where having to turn down overdraft applicants had a negative impact on their perception of service usability. Considering these facts, the Welcome location is judged likely to be unsuitable for most SID offers.

The two competing locations – ID&V and Transaction – both occur after the caller has been identified and it is therefore possible to make the SID offer highly personalised and also to keep track of which customers have previously been approached. The ID&V location has the advantage of being more immediate, while the SID offer in the Transaction location can be tailored more to the result of selecting a service option (e.g. balance information). A potential problem with the Transaction location – and with other attempts to locate the SID offer after the caller has completed the primary purpose (tasks) of the phone call – is that the caller might hang up immediately after obtaining the account information and could miss out on hearing the offer altogether. Therefore, SID offers that need more prompt attention are more suitably located immediately after ID&V.

It is of course possible to envisage alternative solutions to delivering SID offers in the dialogue that take advantage of more than one of the contrasting locations defined above. For example, user behaviour (e.g. how the user operates the automated service) could be logged and then used for a more strategic SID delivery; this approach could, for example, be used to deploy a SID offer immediately after ID&V for those callers who repeatedly tend to hang up before hearing the system "goodbye" message.

8.2.3 Register

Users of automated telephone services can expect a predictable interaction where the dialogue, turn-taking and prompts generally do not change between phone calls to the service. SID offers constitute interruptions in that they suspend the usual dialogue turn-taking for as long as it takes to deliver the new message to the caller (in the experiments there were also no means by which the participants could stop the SID offer). Because callers are unlikely to expect the system to initiate a digression, the opening phrase needs be prominent enough to capture callers' attention. The register employed in these interruptions has significant bearing on how the SID offer is received – as confirmed by the current research.

The approach taken in Experiments 1-3 was to adopt a register that was designed to blend in with the existing system prompts. The opening phrase in the SID offer was "you might like to know that...". In Experiment 4, a more forceful approach was employed by starting the message with an explicit interruption; this was predicted to be perceived as particularly intrusive by participants and, in order to explore whether or not this negative impact could be mitigated, politeness strategies (Positive, Negative and Bald registers) were employed deriving from established face-redress theories in human-human communication.

In fact, the opening phrase in Experiments 1-3 "you might like to know", with its hedging "might" expression, is an example of a 'Negative' politeness strategy. It would have been possible to start the SID offer with simply "you can have an overdraft...", however, the hedging phrase marks the start of a new message, softens the intrusion and has the function of giving the caller a couple of seconds to focus their attention to the ensuing information. A further example of a functional expression in dialogues is "thank you" used after the system receives a response to a request from the caller: this expression not only conveys politeness but also gives feedback to the listener that a response has been recognised. There was some evidence in Experiment 4 that the abrupt Bald SID offer strategy, which lacked this sort of mitigating behaviour, may have had a detrimental effect on user concentration resulting in lower task completion.

The hedging phrase "you might like to know" was well received by participants. In contrast, starting a SID offer with an interruption has a detrimental effect on user attitudes – irrespective of the politeness registers employed in an attempt to mitigate the effect – as the results from Experiment 4 showed. All three SID register variants had a consistent and negative impact on user attitudes to the usability of the automated service. Furthermore, after first exposure to a SID offer, there were no significant differences between participant attitude responses to the three politeness registers. The analysis of the pooled response data (after participants had experience each of the three proposals) revealed that there were significant differences overall between the Bald and the Positive face-redress strategies (in favour of the Bald strategy). The results provide some guidance on the design issues involved in attempts to add such digressions to automated telephone dialogues by eliciting participants' preferences for the wordings of such proposals. The Bald strategy received significantly more positive responses in terms of being shorter, less long-winded and contained more relevant information. The Positive face-redress, on the other hand, was found to be significantly more manipulative, patronising and intrusive.

In the post-experiment listening tests, support of the Bald proposal strategy was strengthened: 54% of participants expressed a preference for the Bald strategy with the main arguments that it was shorter and more to the point than the other designs. Interestingly, however, 50% of participants in the (No-proposal) control group chose the Negative face-redress as their preferred proposal strategy. When heard in isolation, the Negative face-redress might have been perceived as the most appropriate design choice when approaching a customer. In the context of the automated telephone banking service, however, the Negative face-redress approach was shown to be judged as lengthy, long-winded, and was perceived to be too apologetic and formal.

8.2.4 Anthropomorphising the spoken dialogue systems

Much of the research of anthropomorphic computer behaviour in human-computer interaction to date has primarily focussed on the visual user interface; the impact of social phenomena,. such as politeness, in the audio-only interface has yet to be fully explored. The current research contributes to the debate on anthropomorphism in computer systems by exploring the issue of endowing a speech-only human-computer dialogue with specific forms of politeness. In contrast to the visual user interface, the audio-only interface is incapable of displaying multiple pieces of information simultaneously; the system will dominate the dialogue for as long as it takes to deliver a spoken message and the user is not offered the opportunity to rapidly scan information that seems irrelevant. It follows that choice of appropriate wording, duration and speaker characteristics are pivotal in the design of audio interfaces – as demonstrated in this research.

For a given communicative situation between humans, it has been shown that the choice of politeness strategy depends on the mutual expectations about the power relationship and social distance between the interlocutors, coupled with the degree of imposition involved in making the face-threatening act in that communicative context. When a speaker is overly polite, unexpectedly unfriendly or irrational, or strays from the topic in a human-human conversation, the addressee will draw conclusions about the reasons why the speaker does not behave as expected. This may, for example, involve re-evaluating the assumptions about their social relationship with the consequence that politeness (or its absence) in a dialogue can serve to modify the social distance or power relationship between interlocutors. In the case of human-computer interaction, the relationship between the user and the machine is further impacted by the 'black-sheep' effect: social errors are more consequential and critically evaluated if the user knows that the conversational partner is a computer.

Previous research has shown that users are attracted to agent characteristics that are similar to their own personality (submissive/dominant) and are sensitive to consistency in agent personality (introvert/extrovert). The negative reactions towards the face-redressive strategies employed in the proposals in the current research may be attributed to the fact that these were not perceived as being fully integrated with users' assumptions about the relationship with the service, formed by the speaker characteristics presented in the rest of the banking dialogue. Whilst applications, such as the automated banking service explored in this research, are primarily viewed as tools, with repeat use there is the potential that customers will develop aspects of rapport with the service. Endowing audio-only interfaces with personas, which consistently exhibit Negative or Positive face-redressive behaviour as presented here, may thereby serve to enhance this human-computer relationship. The issues raised here lend themselves to further research in order to obtain a deeper understanding of pronounced forms of speaker characteristics, linguistic behaviour and user expectations (e.g. face wants) unique to voice-only computer interfaces. Such research should establish a reference point for the evaluation by assessing participants' own use of politeness strategies (e.g. through a questionnaire) and capturing their anticipated linguistic behaviour for the computer dialogue.

8.2.5 Hidden menus

The target application in the current research was speech-enabled with a main menu listing; the approach taken was to introduce a hidden overdraft option in the main menu which was not listed among the core service options but could be triggered by using they keyword "overdraft". Participants (Experiment 2) were then instructed (after hearing the SID offer in the preceding phone call) to contact the service and apply for an overdraft on their account. A significant number of participants failed to complete the overdraft request.

Unsurprisingly, the main reason for failing the request appeared to be due to the fact that participants expected to have to *select* the overdraft service option from the main menu. This conclusion was verified through the follow-up Experiment 3 in which the overdraft was listed as one of the options in the main menu and which resulted in all participants being able to complete an overdraft request. Using this approach, the purpose of the SID offer would then be reduced to notifying callers that a new service option has been added, or advertising particular features pertaining to the new option that may be of interest to the caller.

However, adding service options to the main menu in this way does not solve the problem entirely: menu listings would be rendered longer and more cluttered, and any key presses associated with service options would have to be updated. For SID offers to function as a 'bolt on' method for introducing new options into the automated service, alternative strategies would have to be sought. One solution would be to keep the overdraft option hidden and instead revisit the contents and wording of the SID offer. The proposal could aim to resolve the primary cause for participants' failing to obtain an overdraft, i.e. redress their erroneous assumption that only service options listed in the main menu can be selected. In this way, the proposal would be used to 'educate' users that – although the overdraft option is not explicitly listed in the main menu – it can still be accessed by saying the "overdraft" keyword.

An alternative solution would be to modify the core dialogue. If, after playing all main menu service options available the caller remains silent, the system could respond with a more open-ended "or just say which [other] service you are interested in ..." in order to try to encourage hesitant callers to volunteer input rather than select it from the menu of 'core' options. Such a generic sentence would avoid additions or changes to the main menu; however, callers cannot expect to know how speech recognition technology works: if they say "overdraft" and it is rejected by the system they will not know whether this is due to the fact that the machine just did not hear them or because the overdraft service option is, in fact, not there. Further silences or rejected inputs at this stage in the dialogue are probably best handled by offering the option for the caller to be transferred to a human advisor. The generic sentence proposed above could alternatively be replaced with "or, *say* one of the following [overdraft, ...]", listing further options. Using this method, core service options (with associated keypad-button presses) could be preserved while supporting the interaction by providing the user with a list to select from – a list activated by voice only which could reasonably easy be changed and updated.

In speech-enabled applications, menus may facilitate the interaction for novice or infrequent users by promoting a step-by-step interaction, but can also render the interaction unnecessarily stilted and long. The challenge to designers of such applications is to strike a balance between restricting the user inputs and at the same time conveying to the user a conceptual model which allows them to fully exploit the strength and flexibility of the speech recognition technology. A review of the menu structure was outside the scope of the current research but it offers significant research opportunities, as these types of menu solutions are pivotal to mass-marked automated telephone services and have significant bearing on the user experience of such services. The dialogue engineer must define appropriate menu solutions to be implemented, but also consider how new or less frequently requested service options may be integrated with the existing dialogue once the automated service has been deployed.

A further issue which arose in the hidden menu experiment was that different users employed different strategies when using the service (volunteering/not volunteering input) and that gender differences appeared to make participants differently equipped in overcoming the problem of the missing menu option. The underlying factors, whether due to differences in mental models of the service functionality or caused by variations in cognitive ability, were not fully addressed in the experiment. It is suggested that future studies into hidden menu options should also probe the mental models amongst participants who *succeeded* in obtaining the hidden option to complement the data from participants who failed to do so. How did these participants reason? Were participants, who volunteered input rather than selecting it from the menu, aware that they did not always wait to hear the option first? What exactly prompted them to request the overdraft?

Furthermore, the demographic data captured in the current experiment was limited; more detailed information about the type of telephone applications participants have experienced in real-life and the frequency of use would need to be obtained in order to provide a full account of the potential impact that habituation effects may have on users' mental models. Such data could then be used to explore if there is a link between habituation effects and user strategies for operating the automated services (e.g. the tendency to volunteer input or select options from menus).

Thirdly, and perhaps most prominently, are questions raised by the gender differences revealed in the experiments. In order to understand the full significance of the impact of gender differences on the usability of automated services, and in order to establish whether or not there is a link between cognitive ability and task completion, further research is necessary. There are standardised psychometric tools available that can be used to measure differences in cognitive abilities (e.g. spatial ability, verbal ability, problem solving skills) and which have been applied in previous research into automated telephone services showing that cognitive ability has a significant impact on service usability and user attitudes. For example, Foster et al. 1998 noted that the level of user's spatial ability affected how participants rated the mode of data entry (touch-tone buttons, isolated and connected word recognition for speech input); level of spatial and verbal ability has also been found to correlate significantly with participant's performance when operating telephone services (CCIR Report on Panel Experiment 5 1995). Goldstein et al. 1999 found that task completion times varied according to spatial ability and prompt strategy: a guided (menu) prompt strategy seemed to better suit participants with low spatial ability and the open ("what do you want to do now") prompt strategy was more suitable for participants with high spatial ability. Other factors such as ageing will also have an impact on users' cognitive ability, affecting their ability to operate automated telephone services (Dulude 2002).

Research into differences in cognitive ability between men and women has met with some controversy (raising issues whether the differences are due to biological or social factors) and the significance of such findings has often been disputed. However, it is generally accepted that men, on average, are better at a range of spatial skills than are women; whereas women are better at some tasks requiring memory abilities for the location of objects (Kimura 1999). Evaluation of participants' cognitive skills was outside the scope of the experiments. It is suggested that future experiments, which aim to explore participants' ability to navigate through automated telephone services, also include an element of psychometric evaluation. Ultimately, in the mass-market domain of self-service applications, the aim is to devise one single dialogue version that strikes a balance between several user groups with different background knowledge, cognitive styles and abilities.

8.3 Final remarks

The dialogue prompts and service functionality have a major impact on how a SID will be received by users. The dialogue featured in the current research was a system-driven, prompt/response telephone application where the core dialogue does not change between phone calls. In contrast, mixed-initiative systems are now beginning to emerge which allow the user to respond to more open prompts – to "say anything". This alters the balance of power by allowing the user to initiate turn-taking. Furthermore, such a system is less reliant on directive prompting and is likely to function better with the deployment of SIDs (such as information about new services, hints or suggestions) since it offers a more flexible dialogue where there are no menu structures that need to be updated and maintained.

There is opportunity for further research in this area, as each new system-initiated digression design variant needs to be assessed in terms of its impact on service usability, user attitudes and task completion. The current research has also highlighted a demand for further studies into issues surrounding the social characteristics conveyed through system prompts and users' abilities to navigate through human-computer dialogues. With automated telephone services becoming ever more ubiquitous in society, and with the increased application of speech recognition in such services, this is a research domain well worth exploring.

The usability evaluations presented here serve to support the thesis that system-initiated digressions can be included as a means for introducing new products and services into the existing dialogue of a speech-enabled automated telephone service.

Bibliography

ALLEN, J. F., BYRON, D.K., DZIKOVSKA, M., FERGUSON, G., GALSESCU, L., STENT, A. (2001). "Towards Conversational Human-Computer Interaction", AI Magazine, Vol. 22, No. 4, American Association for Artificial Intelligence, 2001, pp. 27-37.

ALWIN, D.F., (1992). "Information transmission in the survey interview: number of response categories and the reliability of attitude measurement", Sociological Methodology, Vol. 22 (1992), pp. 83-118.

ANDERSON R.I. (ED.), (2000). "Conversations with Clement Mok and Jakob Nielsen, and with Bill Buxton and Clifford Nass", Interactions, vol. 7 no. 1, Jan.-Feb. 2000, pp. 46-80.

ANDRÉ, E., REHM, M., MINKER, W., BÜHLER., D., (2004). "Endowing spoken language dialogue systems with emotional intelligence", Lecture Notes in Computer Science Vol. 3068, Proceedings of Affective Dialogue Systems: Tutorial and Research Workshop, ADS 2004, Kloster Irsee, Germany, June 14-16, 2004.

ARONS, B., (1991). "Hyperspeech: Navigating in speech-only hypermedia". In Proceedings of Hypertext (San Antonio, TX, Dec. 15-18), ACM, New York, 1991, pp. 133-146.

BAILEY, B.P., KONSTAN, J.A., CARLIS, J.V.(2001). "The Effects of Interruptions on Task Performance, Annoyance, and Anxiety in the User Interface", *Proceedings of the IFIP TC.13 International Conference on Human-Computer Interaction*, Tokyo, Japan, 2001, pp. 593-601.

BERNSEN, N.O., DYBKJÆR, H., DYBKJÆR, L., (1996). "Principles for the design of cooperative spoken human-machine dialogue". Proceedings of the International Conference on Spoken Language Processing (ICSLP) '96, Philadelphia, October 1996, 729-732.

BERNSEN, N.O., DYBKJÆR, H., DYBKJÆR, L., ZINKEVICIUS, V., (1997a). "Generality and transferability. Two issues in putting a dialogue evaluation tool into practical use", Proceedings of Eurospeech'97.

BERNSEN, N.O., DYBKJÆR, H., DYBKJÆR, L., (1997b). "What should your speech system say?", IEEE Computing Practices, December 1997, pp. 25-31.

BEVAN, N., AND MACLEOD, M., (1994), "Usability measurement in context", Behaviour and Information Technology, 13, pp. 132-145.

BOND, C., CAMACK, M., (1999). "Your call is important to us... Please hold", Ergonomics in Design, October 1999, pp. 9-15.

BONITO, J.A., BURGOON, J.K., BENGTSSON, B., (1999). "The role of expectations in human-computer interaction", Proceedings of the International ACM SIGGROUP conference on Supporting group Work, Phoenix, Arizona, US, pp. 229-238.

BOX, G.E.P., (1953). "Non-normality and tests on variances", Biometrika, Vol. 40, No. 3/4 (Dec., 1953), pp. 318-335.

BOYCE, S. J., (2000). "Natural Spoken Dialogue systems for Telephony Applications.", Communications of the ACM, September 2000, Vol. 43, No. 9, pp. 29-34.

BRENNAN, S.E., OHAERI, J., (1994). "Effects of Message Style on User's Attributions towards Agents", Conference Companion CHI'94, April 24-28, 1994.

BROWN, P., LEVINSON, S.C. (1987). "Politeness: some universals in language usage", *Cambridge University Press*, ISBN 0-521-31355-4.

BURGOON, J. K., BENGTSSON, B., BONITO, J., RAMIREZ, JR., A., DUNBAR, N. E. (1999). "Designing Interfaces to Maximize the Quality of Collaborative Work". In R. H. Sprague, Jr. (Ed.), Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences. 1999. HICSS-32.

CARPENTER, B., CHU-CARROLL, J., (1998). "Natural language call routing: a robust self-organizing approach.", Proceedings of ICSLP'98, Sydney, Australia.

CASSELL, J., BICKMORE, T., (2000). "External Manifestations of Trustworthiness in the Interface". In Communications of the ACM. December 2000 Vol. 43(12).

CASSELL, J., BICKMORE, T., (2003). "Negotiated collusion: modelling social language and its relationship effects in intelligent agents", User Modeling and User-Adapted Interaction 13: pp. 89-132.

CHEN, R., (2001). "Self-politeness: A proposal". Journal of Pragmatics 33 (2001). pp. 87-106.

CCIR REPORT ON PANEL EXPERIMENT 5, (1995). "Individual differences and navigation through complex dialogues", Intelligent Dialogues for Automated Telephone Services.

COCHRAN, W.G., (1947). "Some consequences when the assumptions for the analysis of variance are not satisfied", Biometrics, Vol. 3, No. 1 (Mar., 1947), pp. 22-38.

COHEN, M.H., GIANGOLA, J.P., BALOGH, J., (2004). "Voice User Interface Design", *Addison-Wesley*, ISBN 0-321-18576-5.

COLÓN, J. X. E., PÉREZ-QUIÑONES, M. A., FERREIRA, R. (2001): "Effects of Face-Threatening Acts in Human-Computer Dialogues". Proceedings of HFES.

COOLICAN, H., (2004). "Research methods and statistics in psychology", Fourth edition, ISBN 0340812583, Hodder Arnold H&S Publisher.

CULPEPER, J., (1996). "Towards an anatomy of impoliteness". Journal of Pragmatics, 25 (1996). pp. 349-367.

CUTRELL, E., CZERWINSKI, M., HORVITZ, E. (2001). "Notification, Disruption, and Memory: Effects of Messaging Interruptions on Memory and Performance". Proceedings of Interact 2001: IFIP Conference on Human-Computer Interaction, Tokyo, Japan, July 2001.

CZERWINSKI, M., CUTRELL, E., HORVITZ, E., (2000). "Instant Messaging: Effects of Relevance and Time". In S. Turner, P. Turner (Eds), People and Computers XIV: Proceedings of HCI 2000, Vol. 2, September 2000, British Computer Society, p. 71-76.

DIALOGUES 2000 REPORT, (1997). "Navigation in structured and unstructured menus", Experiment Series Report No. 6, Centre for Communication Interface Research, University of Edinburgh of Edinburgh.

DIX, A., FINLAY, J., ABOWD, G.D., BEALE, R., (2004). "Human-Computer Interaction", *Pearson Education Ltd. 2004,* ISBN 0130-461091.

DULUDE, L., (2002). "Automated telephone answering systems and aging", Behaviour & Information Technology, 2002, Vol.21, No. 3, pp. 171-184.

DUTTON, R.T., FOSTER, J., JACK, M.A., STENTIFORD, F.W., (1993). "Identifying usability attributes of automated telephone services", In Proceedings of Eurospeech'93, pp.1335-1338, September 1993.

DUTTON, R.T., FOSTER, J.C., JACK, M.A., (1999). "Please mind the doors – do interface metaphors improve the usability of voice response services?", BT Technology Journal, Vol. 17, No. 1, January 1999.

DYBKJÆR, L., BERNSEN, N.O., DYBKJÆR, H., (1996). "Grice Incorporated: cooperativity in spoken dialogue", Proceedings of the 16th Conference on Computational Linguistics – Vol. 1, August 1996.

DYBJKÆR, L., BERNSEN, N.O., (1998). "A methodology for diagnostic evaluation of spoken human-machine dialogue", International Journal of Human-Computer Studies 48, pp. 605-625.

DYBKJÆR, L., BERNSEN, N. O., (2000). "Usability issues in spoken dialogue systems", Natural Language Engineering 6 (3-4), pp. 243-271.

DYBKJÆR, L. BERNSEN, N. O., (2001). "Usability evaluation in spoken language dialogue systems", In Paroubek, P. and Novick, D. G. (Eds.): Proceedings of the Workshop on Evaluation Methodologies for Language and Dialogue Systems, Association for Computational Linguistics 39th Annual Meeting and 10th Conference of the European Chapter (ACL/EACL) 2001, Toulouse, France, 6-7 July 2001, pp. 9-18.

DYBKJÆR, L., BERNSEN, N. O., AND MINKER, W., (2004). "New challenges in usability evaluation – beyond task-oriented spoken dialogue systems", Proceedings of The International Conference for Spoken Language Processing, ICSLP 2004, South Korea, 2004.

EDWARDS, K., QUINN, K., DALZIEL, P.B., JACK, M.A., (1997). "Evaluating commercial speech recognition and DTMF technology for automated telephone banking services", IEE Colloquium on Advances in Interactive Voice Technologies for Telecommunications Services, IEE Digest No. 1997/147, pp. 1-6, June 1997.

ESCANDELL-VIDAL, V. (1996). "Towards a cognitive approach to politeness". Language Sciences, Vol. 18, pp. 629-650. 1996.

ETSI ETR 095. "Human factors (HF); Guide for usability evaluations of telecommunications systems and services", European Telecommunications Standards Institute (ETSI), Technical Report DTR/HF-3001, September 1993.

ETSI ETR 096. "Human factors (HF); Phone based interfaces (PBI) human factors guidelines for the design of minimum phone based user interface to computer services", European Telecommunications Standards Institute (ETSI), Technical Report DTR/HF-1002, August 1993.

ETSI ETR 329. "Human factors (HF); Guidelines for procedures and announcements in Stored Voice Services (SVS) and Universal Personal Telecommunication (UPT)", European Telecommunications Standards Institute (ETSI), Technical Report DTR/HF-01029, December 1996.

ETSI EG 201 472. "Human factors (HF); Usability evaluation for the design of telecommunication systems, services and terminals", European Telecommunications Standards Institute (ETSI), Technical Report DEG/HF-00006, September 1993.

ETSI EG 202 076. "Human factors (HF); User Interfaces; Generic spoken command vocabulary for ICT devices and services", European Telecommunications Standards Institute (ETSI), Technical Report DEG/HF-00055, November 2002.

ETSI EG 202 116. "Human factors (HF); Guidelines for ICT products and services; 'Design for all'", European Telecommunications Standards Institute (ETSI), September 2002.

FIELD, A., (2000). "Discovering statistics using SPSS for Windows: Advanced techniques for beginners", SAGE Publications Ltd, ISBN 0761957553.

FIELD, A. & HOLE, G., (2003). "How to design and report experiments", SAGE Publications Ltd, ISBN 0-7619-7383-4.

FISCHER, G. (2001). "User Modeling in Human-Computer Interaction", Contribution to the 10th Anniversary Issue of the Journal "User Modeling and User-Adapted Interaction (UMUAI)", Vol. 11, No. 1/2, pp 65-86, 2001.

FOGG, B.J., NASS C. (1997). "Silicon sycophants: the effects of computers that flatter", International Journal of Human-Computer Studies (1997) 46, pp. 551-561.

FOSTER, J.C., DUTTON, R., JACK, M.A., LOVE, S., NAIRN, I.A., VERGEYNST, N., STENTIFORD, F.W.M., (1992). "Design and evaluation of dialogues for automated telephone services", Proc. Institute of Acoustics, vol.14, no.6, pp.629-635, November 1992.

FOSTER, J.C., MCINNES, F.R., JACK, S., LOVE, S., DUTTON, R.T., NAIRN, I.A., WHITE, L.S. (1998). "An experimental evaluation of preferences for data entry method in automated telephone services", Behaviour & Information Technology, 1998, Vol. 17, No. 2, pp.82-92.

FRANKE, J.L., DANIELS, J.J., MCFARLANE, D.C., (2002). "Recovering context after interruption", In 24th Annual Meeting of the Cognitive Science Society. CogSci. Fairfax, VA.

FRØKJÆR, E., HERTZUM, M., HORNBÆK, K., (2000). “Measuring usability: are effectiveness, efficiency and satisfaction really correlated”CHI Letters, Vol. 2, Issue 1, 1-6 April 2000, pp. 345-352.

FURMAN, D.S., COSKY, M.J., THOMSON, D.L., O’BRIEN, S.A., SUMNER, E.E., (1999). “Speech-based services”, Bell Labs Technical Journal, April-June 1999, pp. 88-97.

GARDNER-BONNEAU, D. (ED.) (1999). “Human Factors and Voice Interactive Systems”, *Kluwer Academic Publishers, Second Printing 2001*, ISBN 0-7923-8467-9.

GLASS, J.R., (1999). “Challenges for spoken dialogue systems”, In Proceedings of the 1999 IEEE ASRU Workshop.

GOLDSTEIN, M., BRETAN, I., SALLNÄS, E.-L., BJÖRK, H., (1999). “Navigational Abilities in Audial Voice-Controlled Dialogue Structures”, In Behaviour & Information Technology, 1999, Vol. 18, No. 2, pp. 83-95.

GOLDSTEIN, M., ALSIÖ, G., WERDENHOFF, J., (2002). “The media equation does not always apply: People are not polite towards small computers”, Personal and Ubiquitous Computing, Vol. 6, Issue 2 (April 2002), pp. 87-96.

GORIN, A.L., RICCARDI, G., WRIGHT, J.H. (1997). “How may I help you?”. Speech Communication, Vol.23 (1997), pp.113-127.

GREEN, N., CARBERRY, S., (1999). “A Computational Mechanism for Initiative in Answer Generation”. User Modelling and User-Adapted Interaction, 9 (1/2), April 1999, pp. 93-132.

GREEN, A., (2002). “Usability design for spoken language dialogue”, http://www.ida.liu.se/~nlplab/gslt/papers/AndersG_final2.pdf (availability last checked 13/06/05).

GRICE, H.P., (1975). "Logic and conversation", In Syntax and Semantics, Cole, P., and Morgan, J.L. (Eds.), Vol 3. pp. 41-58, Academic Press, Inc., ISBN 0-12-785423-1.

GRUDIN, J., (1989). "The case against user interface consistency", Communications of the ACM, October 1989, Volume 32, Number 10, pp 1164-1173.

GUILFORD, J.P. (1967). "Response biases and response sets", In Readings in Attitude Theory and Measurement, Fishbein, M., (ed.), John Wiley & Sons. Inc., pp. 277-281.

GUSTAFSON, J., BELL, L., (2000). "Speech technology on trial: Experiences from the August system", Natural Language Engineering 6 (3-4), pp. 273-286.

GUTTMAN, L., (1977). "What is not what in statistics", The Statistician, Vol. 26, No. 2 (jun., 1977), pp. 81-107.

HALLER, S., (1994). "Recognizing Digressive Questions", In Proceedings of AAAI94, Fall Symposium 1994.

HALLER, S., S. MCROY (1997). "Computational Models for Mixed Initiative interactions", Papers from the 1997 AAAI Spring Symposium. Technical Report: SS-97-04: AAAI/MIT Press, 1997.

HALSTEAD-NUSSLOCH, R., (1989). "The Design of Phone-Based Interfaces for Consumers", CHI'89 Proceedings, May 1989, pp. 347-352.

HANSEN, B., NOVICK, D. G., SUTTON, S., (1996). "Systematic Design of Spoken Prompts", Proceedings of the SIGCHI conference on Human factors in computing systems: common ground, Vancouver, British Colombia, Canda, 1996, pp. 157-164.

HARTIKAINEN, M., SALONEN, E-P., TURUNEN, M., (2004). "Subjective evaluation of spoken dialogue systems using SERVQUAL method", Proc. of ICSLP'04.

HEARST, M. A. ALLEN, J. F., C. I. GUINN, E. HORVITZ (1999). "Mixed-initiative interaction", IEEE Intelligent Systems, Vol. 14(5): pp. 14-23, Sept-Oct 1999.

HONE, K.S., BABER, C. (1999). "Modelling the effects of constraint upon speech-based human–computer interaction", International Journal of Human-Computer Studies, January 1999, vol. 50, no. 1, pp. 85-107.

HONE, K. S., GRAHAM, R., (2000). "Towards a tool for the Subjective Assessment of Speech System Interfaces (SASSI)", Natural Language Engineering 6 (3-4), pp.287-303.

HONE, K. S., BABER, C. (2001). "Designing habitable dialogues for speech-based interaction with computers". Int. J. Human-Computer Studies (2001) 54, pp. 637-662.

HONE, K. S., GRAHAM, R., (2001). "Subjective assessment of speech-system interface usability", Proceedings of Eurospeech , Scandinavia 2001.

HORNSTEIN, T. (1994). "Telephone Voice Interfaces on the Cheap", In Bischofberger, W.R., Frei, H.P. (eds), *Computer Science Research at UBILAB, Strategy and Projects, Proceedings of the UBILAB '94 Conference, Zurich*, pp. 134-146, Universitätsverlag Konstanz, Konstanz, September 1994.

HORVITZ, E., (1999). "Principles of Mixed-Initiative User Interfaces". In Proceedings of ACM CHI 99 Conference on Human Factors in Computing Systems, volume 1 of Characters and Agents, pages 159-166, 1999.

ISBISTER, K., NASS, C., (2000). "Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics", International Journal of Human-Computer Studies, 53, pp. 251-267.

ISO 9241-11, (1998). "Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11: Guidance on usability".

JACK, M.A., FOSTER, J.C., STENTIFORD, F.W., (1992a). "Design and evaluation of automated telephone services", IEE Colloquium on Telecommunications, Consumer and Industrial Applications of Speech Technology, IEE Digest No. 1992/144, pp. 611-614, 1992.

JACK, M.A., FOSTER, J.C., STENTIFORD, F.W., (1992b). "Intelligent dialogues in automated telephone services", In Proceedings of the International Conference on Spoken Language Processing (ICSLP'92), Banff, Alberta, Canada, October 1992.

JACK, M.A., FOSTER, J.C., STENTIFORD, F.W., (1993). "Usability analysis of intelligent dialogues for automated telephone services", Proc. Joint ESCA/NATO workshop on Applications of Speech Technology, pp.149-152, September 1993.

JAMIESON, S., (2004). "Likert scales: how to (ab)use them", Commentary in Medical Education 2004, Vol. 38, pp.217-218.

JOHNSON, W.L., RIZZO, P., BOSMA, W., KOLE, S., GHIJSEN, M., VAN WELBERGEN, H., (2004). "Generating socially appropriate tutorial dialog", Lecture Notes in Computer Science, Vol. 3068, May 2004, pp. 254-264.

JOHNSON, W.L., RIZZO, P., (2004). "Politeness in tutoring dialogs: run the factory that's what I'd do", Lecture Notes in Computer Science, Vol. 3220, pp. 67-76.

JONES, E. E., (1986). "Interpreting interpersonal behaviour: The effects of expectancies", Science, New Series, Vol. 234, No. 4772 (Oct. 3, 1986), pp 41-46.

KAMM, C. A., LITMAN D. J., WALKER, M. A. (1998). "From novice to expert: the effect of tutorials on user expertise with spoken dialogue systems", Proceeding of the International Conference on Spoken Language Processing (ICSLP98).

KARIS, D., DOBROTH, K.M., (1991). "Automating Services with Speech Recognition over the Public Switched Telephone Network: Human Factors Considerations.", IEEE Journal on Selected Areas in Communication, Vol. 9, No. 4, May 1991.

KARSENTY, L., (2001). "Adapting verbal protocol methods to investigate speech systems use", Applied Ergonomics 32, pp. 15-22.

KIMURA, D., (1999). "Sex and Cognition", MIT Press, 1999, ISBN 0-262-11236-1.

KOTELLY, B., (2003). "The art and business of speech recognition: creating the noble voice", Addison-Wesley, ISBN 0-321-15492-4.

KNAPP, T.R., (1990). "Treating ordinal scales as interval scales: an attempt to resolve the controversy", Nursing Research, March/April 1990, Vol. 39, No. 92, pp. 121-123.

KRAHMER, E., LANDSBERGEN, J., POUTEAU, X., (1997) "How to Obey the 7 Commandments for Spoken Dialogue Systems". In: Proceedings of the (E)ACL workshop on Interactive Spoken Dialog Systems, J. Hischberg, C. Kamm & M. Walker (eds.), Madrid, 82-89, 1997.

KRÜGER, H., KRUCKENBERG, H., (1999). "Dialogue design in phone-based interfaces", The 17th International Symposium on Human Factors in Telecommunication, May 1999, Copenhagen, Denmark.

LABOVITZ, S., (1967). "Some observations on measurement and statitstics", Social Forces, Vol. 46, No. 2 (Dec., 1967), pp. 151-160.

LAI, J., (2000). "Conversational Interfaces". Communications of the ACM, September 2000, Vol. 43, No. 9, pp. 24-27.

LAMEL, L., ROSSET, S., GAUVAIN, J., (2000). "Consideration in the design and evaluation of spoken language dialogue system", Proceedings of ICSLP'00, 6th International Conference on Spoken Language Processing, Beijing China, October 2000, Vol. 4, pp.5-8.

LARSEN, B.L. (1997a). "A Strategy for Mixed-initiative Dialogue Control", Proceedings of Eurospeech '97, September, 1997.

LARSEN, L.B., (1997b). "Investigating a mixed-initiative dialogue management strategy", In Proceedings of the 1997 IEE Workshop on Automatic Speech Recognition and Understanding, December 1997,
Santa Barbara, CA , USA, pp. 65-71.

LARSEN, L.B., (2003). "Assessment of spoken dialogue system usability – what are we really measuring?", Eurospeech'03, pp. 1945-1948.

LEE, C-H., CARPENTER, B., CHOU, W., CHU-CARROLL, J., REICHL, W., SAAD, A., ZHOU, Q., (2000). "On natural language call routing", Speech Communication 31, pp. 309-320.

LEE, K.M., NASS, C., (2003). "Designing social presence of social actors in human computer interaction", CHI Letters, Volume No. 5, Issue No. 1.

LEECH, G.N. (1983). "Principles of Pragmatics", Longman Group Ltd., ISBN 0-582-55110-2.

LENK, U., (1998). "Discourse markers and global coherence in conversation". Journal of Pragmatics 30 (1998) pp. 245-257.

LEWIS, J.R., (2002). "Psychometric evaluation of the PSSUQ using data from five years of usability studies", International Journal of Human-Computer Interaction, 14(3&4), pp. 463-488.

LIKERT, R. (1932). "A Technique for the Measurement of Attitudes", Archives of Psychology, Vol. 140, June 1932, pp. 5-55.

LIKERT, R., (1967). "The method of constructing an attitude scale", In Readings in Attitude Theory and Measurement, Fishbein, M., (ed.), John Wiley & Sons. Inc., pp. 90-95.

LOVE, S., DUTTON, R.T., FOSTER, J.C., JACK, M.A., NAIRN, I.A., VERGEYNST, N.A., STENTIFORD, F.W.M., (1992). "Towards a usability measure for automated telephone services", Proc. Institute of Acoustics Speech and Hearing Workshop, vol.14, no.6, pp.553-559, November 1992.

LOVE, S., DUTTON, R.T., FOSTER, J.C., JACK, M.A., STENTIFORD, F.W.M., (1994). "Identifying salient usability attributes for automated telephone services", Proceedings of the Interational Conference of Speech and Language Processing '94, Yokohama, Japan.

MARAKAS, G.M., JOHNSON, R.D., PALMER, J.W., (2000). "A theoretical model of differential social attributions toward computing technology: when the metaphor becomes the model", International Journal of Human-Computer Studies, 52, pp. 719-750.

MARICS, M., ENGELBECK, G., (1997). "Designing voice menu applications for telephones", Handbook of Human-Computer Interaction, Second completely revised edition, Helander, M., Landauer, T.K., Prabu, P. (eds.), Elsevier Science B.V., ISBN 0444818626.

MCFARLANE, D. (1997), "Interruption of People in Human-Computer Interaction: A General Unifying Definition of Human Interruption and Taxonomy", Technical Report NRL/FR/5510-97-9870, US Naval Research Lab, Washington, DC.

MCFARLANE, D. (1998), "Interruption of People in Human-Computer Interaction, PhD Thesis, The George Washington University, August 1998.

MCFARLANE, D.C., LATORELLA, K.A. (2002) "The scope and importance of human interruption in HCI design". Journal of Human-Computer Interaction, Vol. 17, No.3.

MCINNES, F.R. , NAIRN, I.A., ATTWATER, D.J., JACK, M.A., (1999). "Effects of prompt style on user responses to an automated banking service using word-spotting", BT Technology Journal, vol.17, no.1, pp.160-171, January 1999.

MCTEAR, M.F., (2004). "Spoken dialogue technology: toward the conversational user interface", *Springer-Verlag London Ltd*, ISBN 1-85233-672-2.

MEIER, A. J., (1995a). "Passages of politeness". Journal of Pragmatics. 24 (1995) pp. 381-392.

MEIER, A. J., (1995b). "Defining Politeness: Universality in Appropriateness", Language Sciences. Vol. 17, No. 4. pp. 345-356 (1995).

MILLER C.A., (2004). "Human-computer etiquette: managing expecations with intentional agents", Communications of the ACM, April2004/Vol.47, No. 4, pp. 31-34.

MOLICH, R., NIELSEN, J. (1990). "Improving a Human-Computer Dialogue", Communications of the ACM, March 1990, Vol. 33, Num. 3, pp. 338-348.

MSN BC (1999). "Is Your PC Too Friendly?", Commentary Article in ZD Net News, http://zdnet.com.com/2100-1107-513612.html?legacy=zdnn (link last checked 18/06/05).

NARAYANAN, S., G. DI FABBRIZIO, C. KAMM, J. HUBBELL, B. BUNTSCHUH, P. RUSCITTI, AND J. WRIGHT (2000). "Effects of dialog initiative and multi-modal presentation strategies on large directory information access", In Proc. of ICSLP, (Beijing, China), pp. 636-639, 2000.

NASS, C., GONG, L. (2000). "Speech Interfaces from an Evolutionary Perspective". Communications of the ACM, Vol. 43, No. 9, September 2000.

NASS, C., LEE, K.M., (2001). "Does Computer-Synthesized Speech Manifest Personality? Experimental Test of Recognition, Similarity Attraction, and Consistency-Attraction". Journal of Experimental Psychology:Applied, Vol. 7, No. 3, pp. 171-181.

NASS, C., MOON, Y., (2000). "Machines and Mindlessness: Social Responses to Computers", Journal of Social Issues, Vol. 56, No. 1, 2000, pp. 81-103.

NASS, C., STEUER, J., TAUBER, E.R. (1994). "Computers are social actors", Proceedings of the CHI Conference, CHI'94, Boston, MA., pp. 72-78.

NIELSEN, J. (1993). "Usability engineering", *Academic Press, Inc.,* ISBN 0-12-518405-0.

NIELSEN, J., MACK, R.L. (1994). "Usability inspection methods", John Wiley & Sons, Inc., ISBN 0-471-01877-5.

O'DRISCOLL, J., (1996). "About face: A defence and elaboration of universal dualism". Journal of Pragmatics 25 (1996). pp. 1-32.

OPPENHEIM, A.N. (1992). "Questionnaire Design, Interviewing and Attitude Measurement", Pinter Publications Ltd, ISBN 1855670445.

OULASVIRTA, A., SAARILOUMA, P., (2004). "Long-term working memory and interrupting messages in human-computer interaction", Behaviour & Information Technology, 2004, Vol. 23, No. 1, pp. 53-64.

PALLANT, J. (2001). "SPSS survival manual", Open University Press, Maidenhead, Philadelphia, 2001, ISBN 0-335-20890-8.

PEIRIS, D.R., GREGOR, P., ALM, N. (2000). "The effects of simulating human conversational style in a computer-based interview". Interacting with Computers 12 (2000) pp. 635-650.

PREECE, J., ROGERS, Y., SHARP, H., BENYON, D., HOLLAND, S., CAREY, T., (1994). "Human-Computer Interaction", *Pearson Education Ltd.,* ISBN 0-201-62769-8.

PREECE, J., ROGERS, Y., SHARP (2002). "Interaction Design", *John Wiley & Sons, Inc.* ISBN 0-471-49278-7.

PRESTON, C.C., COLMAN, A.M., (2000). "Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences", Acta Psychologica 104, pp1-15.

RAMAKRISHNAN, N., R. CAPRA, M. A. PÉREZ-QUIÑONES (2002). "Mixed-Initiative Interactio*N* = Mixed Computation", In Proceedings of the ACM SIGPLAN Workshop on Partial evaluation and Semantic-Based Program Manipulation (PEPM'02) (P. Thieman, ED.), pp 119-130, January 2002.

RASKIN, J., (2000). "The humane interface: new directions for designing interactive systems", *Addison-Wesley,* ISBN 0-201-37937-6.

REEVES, B., NASS, C. (1996). "The Media Equation; How People Treat Computers, Television, and New Media Like Real People and Places". Cambridge University Press/CSLI, New York, ISBN 1-57586-052-X.

RESNICK, P., VIRZI, R.A., (1992). "Skip and Scan: Cleaning up Telephone Interfaces", Proceedings of Human Factors in Computing Systems, 1992.

RIBEIRO, N.M. & BENEST, I.D. (2002), "*Invisible but Audible: Enhancing Information Awareness through Anthropomorphic Speech*". In Faulkner, X., Finlay, J. & Detienne, F., (Eds), People and Computers XVI - Memorable Yet Invisible, Proc. of HCI 2002 – 16th British HCI Group Annual Conference on Human-Computer Interaction, London, UK, 2 - 6 September 2002, pp. 17-35.

ROOT, R.W., DRAPER, S., (1983). "Questionnaires as a software evaluation tool", Proceedings of CHI'83, December 1983, pp. 83-87.

ROSENFELD, R., OLSEN D., RUDNICKY, A., (2001). "Universal Speech Interfaces", Interactions November + December 2001.

ROSS, S., BROWNHOLTZ, E., ARMES, R., (2004). "Voice user interface principles for a conversational agent", Proceedings of the 9th international conference on Intelligent user interface, January 2004.

SAYGIN, A. P., CICEKLI, I., (2002) "Pragmatics in human-computer conversations", Journal of Pragmatics 34 (2002), pp. 227-258.

SHEEDER, T., BALOGH, J., (2003). "Say it like you mean it: priming for structure in caller responses to a spoken dialogue system", International Journal of Speech Technology, 6, pp. 103-111.

SHNEIDERMAN, B., (1989). "A Nonanthropomorphic Style Guide: Overcoming the Humpty Dumpty Syndrome", The Computing Teacher 16(7), October 1989.

SHNEIDERMAN, B., (1993). "Beyond Intelligent Machines: Just Do It!", IEEE Software, January 1993, Vol. 10, No. 1, pp. 100-103.

SHNEIDERMAN, B. (1998). "Designing the user interface: strategies for effective human-computer interaction", *Third Edition, Addison-Wesley Publishing Company*, ISBN 0-201-69497-2.

SHNEIDERMAN, B. (2000). "The Limits of Speech Recognition.", Communications of the ACM, September 2000, Vol. 43, No. 9, pp. 63-65.

SPENCER-OATEY, H., (2002). "Managing rapport in talk: Using rapport sensitive incidents to explore the motivational concerns underlying the management of relations". Journal of Pragmatics 34 (2002). pp. 529-545.

STALLARD, D., (2001). "Evaluation results for the Talk'n'Travel system", Proceedings of the First International Conference on Human Lanugage Technology Research.

STEVENS, S.S., (1946). "On the theory of scale measurement", Science, Vol. 103, No. 2684 (Jun. 7, 1946), pp. 677-680.

SUHM, B., PETERSON, P., (2001). "Evaluating commercial touch-tone and speech-enabled telephone voice user interfaces using a single measure", Conference on Human Computer Interaction CHI'01, Interactive Posters, 31 March-5 April.

SUHM, B., BERS, J., MCCARTHY, D., FREEMAN, B., GETTY, D., GODFREY, K., PETERSON, P., (2002). "A comparative study of speech in the call center: Natural language call routing vs. touch-tone menus", CHI Letters, Paper: Speech, Audio, Gesture, Minneapolis, Minnesota, USA, 20-25 April.

TAKEUCHI, Y., KATAGIRI, Y., NASS, C., FOGG, B. J. (1998). "Social Response and Cultural Dependency in Human-Computer Interaction", Proceedings of PRICAI'98, 114-123, 1998.

TAKEUCHI, Y., KATAGIRI, Y., (1999). "Identity Perception of Computers as Social Actors". Proceedings of ICCS/JCSS'99, pp. 922-925.

THURSTONE, L.L., (1967). "Attitudes can be measured", In Readings in Attitude Theory and Measurement, Fishbein, M., (ed.), John Wiley & Sons. Inc., pp.77-89.Johnson, W.L., Rizzo, P., Bosma, W., Kole, S., Ghijsen, M., van Welbergen, H., (2004a). "Generating socially appropriate tutorial dialog", Lecture Notes in Computer Science, Vol. 3068, May 2004, pp. 254-264.

TOMKO, S. AND ROSENFELD, R., (2004) "Shaping Spoken Input in User-Initiative Systems.", Proceedings of ICSLP'04, 8th International Conference on Spoken Language Processing, Jeju, Korea, October 4-8.

TROUVAIN, J., SCHRÖDER, M., (2004). "How (not) to add laughter to synthetic speech", Lecture Notes in Computer Science Vol. 3068, Proceedings of Affective Dialogue Systems: Tutorial and Research Workshop, ADS 2004, Kloster Irsee, Germany, June 14-16, 2004.

TZENG, J., (2004). "Toward a more civilised design: studying the effects of computers that apologise", International Journal of Human-Computer Studies 61, 2004, pp. 319-345.

VANHOUCKE, V., NEELEY, W. L., MORTATI, M., SLOAN, M. J., NASS, C., (2001). "Effects of Prompt Style when Navigating through Structured Data", Proceedings of INTERACT 2001, Eighth IFIP TC.13 Conference on Human Computer Interaction, IOS Press, pp.530--536, Tokyo, Japan, 2001.

VAN KUPPEVELT, J., HEID, U., KAMP, H. (eds.) (2000). "Best practice in spoken language dialogue systems engineering: Introduction to the special issue", Natural Language Engineering 6 (3-4), pp. 205-212.

VELLEMAN, P.F., WILKINSON, L., (1993). "Nominal, ordinal, interval, and ratio typologies are misleading", The American Statistician, Vol. 47, No. 1 (Feb., 1993), pp. 65-72.

WAERN, Y., (1993). "Varieties of Learning to Use Computer Tools", Computers in Human Behavior, Vol.9, pp. 323-339, 1993.

WALKER, M. A., CAHN, J. E., WHITTAKER, S.J., (1997) "Improvising Linguistic Style: Social and Affective Bases for Agent Personality", In Proceedings of the Conference on Autonomous Agents, AGENTS97, 1997.

WALKER, M.A., FROMER, J., DI FABBRIZIO, G., MESTEL, C., HINDLE, D., (1998). "What can I say?: evaluating a spoken language interface to email.", Proc. of ACM CHI 98, Conference on Human Factors in Computing Systems, pp. 582-589.

WALKER, M., KAMM, C., LITMAN, D., (2000). "Towards developing models of usability with PARADISE", Natural Language Engineering 6 (3-4), pp.363-377.

WALKER, M.A., PASSONNEAU, R., BOLAND, J.E., (2001). "Quantitative and Qualitative Evaluation of Darpa Communicator Spoken Dialogue Systems", In *Meeting of the Association of Computational Linguistics* , 2001.

WANG, N., JOHNSON, L., RIZZO, P., SHAW, E., MAYER, R.E., (2005). "Experimental evaluation of polite interaction tactics for pedagogical agents", Proceedings of the 10th international conference on Intelligent user interfaces, San Diego, California, USA , pp. 12 – 19.

WATTS, R. J. (2003). "Politeness", Cambridge University Press, ISBN 0-521-7905.

WEINSCHENK, S., BARKER, D.T., (2000). "Designing effective speech interfaces", *John Wiley & Sons, Inc.,* ISBN 0-471-37545-4.

WHITTAKER, S.J., ATTWATER, D.J. (1996). "Interactive Speech Systems for Telecommunications Applications", BT Technology Journal, Vol. 14, No. 2, April 1996.

WILLIAMS, J.D., SHAW, A., PIANO, L., ABT, M., (2003a). "Evaluating real callers' reactions to Open and Directed Strategy prompts." Applied Voice Input/Output Society Speech Developers Conference/SpeechTEK Spring EXPO, 31 March – 3 April 2003, San Jose, California.

WILLIAMS, J.D., SHAW, A.T., PIANO, L., ABT, M., (2003b). "Preference, perception and task completion of open, menu-based and directed prompts for call routing: a case study", Proceedings of Eurospeech 2003, Geneva.

WILLIAMS, J.D., WITT, S.M., (2004). "A comparison of dialog strategies for call routing", International Journal of Speech Technology 7(1), pp. 9-24, January 2004.

WITT, S.M., WILLIAMS, J.D., (2003). "Two studies of open vs. directed dialogue strategies in spoken dialogue systems", Proceedings of Eurospeech 2003, Geneva, Switzerland.

YANKELOVICH, N., (1995). "Designing SpeechActs: Issues in Speech User Interfaces".In Proceedings of ACM CHI'95, Conference on Human Factors in Computing Systems, Denver, CO: ACM Press, May 7-11, 1995.

YANKELOVICH, N., (1996). "How do users know what to say?", ACM Interactions, Vol. 3, No. 6, November/December 1996, pp32-43.

YU, C. H., (2002). "An overview of remedial tools for violations of parametric test assumptions in the SAS system.", Proceedings of 2002 Western Users of SAS Software Conference, pp. 172-178.Hone, K. S., Graham, R., (2000). "Towards a tool for the Subjective Assessment of Speech System Interfaces (SASSI)", Natural Language Engineering 6 (3-4), pp.287-303.

Appendix 1

Appendix 1.1. – Experiment priming material

Lloyds TSB

PhoneBank *Express* Telephone Banking Service:

Telephone number: (9) 667 1818

Your Customer Details:

Name:	**J Smith**
Membership number:	**255881487**
T.I.N.	
(Telephone Identification Number):	**139272**

Your Account Details:

Current Account:	**0074242**
Savings Account:	**7013818**

Appendix 1.2. – Experiment 1 research script

Research Script – Experiment 1

Thank you for agreeing to take part in this usability experiment. We are going to ask you to try out an automated telephone banking service and also to fill in some questionnaires relating to your opinion of the service.

To start off, I would like to ask you some general questions (**Demographic Questionnaire**).

Before I ask you to phone the service there are some information and details that I have to go through with you. The service we ask you to try today is called 'PhoneBank *Express*'. It is an automated telephone banking service which Lloyds TSB provide their customers.

Here's a copy of the 'mini guide' (**Mini Guide**) which you can use when you phone the service. It contains information about the services available. For example *(pointing at information as you explain)*, you can find out the balance of your account, hear recent transactions or search for a particular transaction in your account by selecting 'item search'. You can either speak your commands to the automated service or press the buttons on the telephone – or use both if you wish. Examples of what you can say and buttons to press are listed here *(show the participant)*.

In order to use PhoneBank *Express*, you will need your personal membership number and an identification number. You get these details sent to you from Lloyds TSB, together with your copy of the mini guide, when you register with to use the service. (**Sheet with Customer Details**)

Today we'd like you to try to imagine that you are 'J Smith' – a Lloyds TSB customer – and that you have registered to use PhoneBank *Express*. These are the details you need to access your banking details through the automated service *(Read out all the details on the sheet, but skip reading out number sequences)*.

I would like you to phone the service in order to try to find out the following *(Hand participant* **Task Sheet 1** *and read out the details)*. Feel free to have a look at the mini guide before you start phoning. When you find out the balance and the information about the cheque, please note down the information on the sheet. Have you got any questions? I'd also like to say that while you make your phone call to the service I will not be able to talk to you and so if you have any questions can you keep these for later? *(participant phones the service)*

Thank you. I'd now like you to imagine that a few days have passed and you call the automated service again to find out the same details as the last time *(hand participant* **Task sheet 2***)*.

-------- *task 3 is carried out in the same manner with 'a few days passing' scenario (**Task Sheet 3**).*---------

Thank you. Now I'd like you to fill out this questionnaire for me please (**Likert Questionnaire 1**). You'll find 20 statements relating to your attitude to the service, each statement worded positively or negatively. For each statement we'd like you to tick the box which best expresses your opinion on that statement – the row of boxes range from strongly agree to strongly disagree. If you have any questions let me know.

Thank you. I would now like you to phone the service again and do the same tasks again (**Task Sheet 4**).

Thank you. Could you fill in this questionnaire again – this time keeping in mind that it relates to your experience and opinion about your most recent phone call to the service. (**Likert Questionnaire 2**)

Thank you. That's all the phone calls you will have to make. I have some final questions I would like to ask you before you leave relating to your opinion about the service. (**Exit Questionnaire**)

Thank you for your participation. *(hand participant £10 cheque)*

Appendix 1.3. – Experiment 1 research procedure

Research Procedure – Experiment 1

1. Meet and greet.
2. Explain the experiment, what the participant will do (phone an automated banking service and fill in questionnaires) and then go through the questions in the **Demographic/Technographic Questionnaire**.
3. Brief information about Lloyds TSB and the PhoneBank *Express* automated telephone banking service.
4. Hand participant **Mini Guide** and go through available services and input modes (explain in more detail about how to get a balance and to find a transaction using 'recent transactions' or 'item search').
5. Explain that you need a personal membership number and TIN to use PhoneBank *Express* and to access the banking details.
6. Hand the customer the sheet with **Customer Details** and go through all the details together, membership number, TIN, account details etc.
7. Give participant **Task Sheet 1** and read through the information, pointing out that we want the participant to note down the information about balance and cheque.
8. Allow for participant to read through details and ask questions before phoning the service. *(participant calls the service)*
9. Explain the scenario that 'a few days have passed' and that the participant now phones up to check the same banking details again, give participant **Task Sheet 2**. *(participant calls the service)*
10. Again, explain that 'a few days have passed' and that the participant now phones up the service again, give participant **Task Sheet 3**. *(participant calls the service)*
11. After the participant completes the third call – hand **Likert Usability Q1** and explain briefly how it is filled out, polarity, 7-point scale from agree to disagree etc.
12. Explain that 'a few days have passed' and that the participant now phones up to check details again, give participant **Task Sheet 4**. *(participant calls the service)*
13. After the participant completes the fourth and final call – hand **Likert Usability Q2**, point out that it is the same questionnaire but that this time it relates to the experience of the most recent call to the service.
14. Finally, explain that that this was the last phone call to the service and go through the questions in the **Exit Questionnaire** – explaining to the participant that we are interested about finding out more details about what he/she thinks about the service.
15. Pay the participant £10, thanking for his/her participation in the experiment.

Appendix 1.4. – Likert Usability (LU) questionnaire

Thinking about the service you have just used. For each statement below, tick the box which best expresses your opinion on that statement.

	Strongly Agree	Agree	Slightly Agree	Neither agree nor disagree	Slightly Disagree	Disagree	Strongly Disagree
Q1 I found the service confusing to use.	□	□	□	□	□	□	□
Q2 When I was usirig the service I always knew what I was expected to do.	□	□	□	□	□	□	□
Q3 I thought the service was efficient.	□	□	□	□	□	□	□
Q4 I felt flustered when using the service.	□	□	□	□	□	□	□
Q5 I would be happy to use the service again.	□	□	□	□	□	□	□
Q6 I would prefer to talk to a human being.	□	□	□	□	□	□	□
Q7 I thought the voice was very clear.	□	□	□	□	□	□	□
Q8 I felt under stress when using the service.	□	□	□	□	□	□	□
Q9 The service was too fast for me.	□	□	□	□	□	□	□
Q10 I liked the voice.	□	□	□	□	□	□	□
Q11 I felt that the service was reliable.	□	□	□	□	□	□	□
Q12 I enjoyed using the service.	□	□	□	□	□	□	□
Q13 The service was friendly.	□	□	□	□	□	□	□
Q14 I think the service needs a lot of improvement.	□	□	□	□	□	□	□
Q15 I thought the service was polite.	□	□	□	□	□	□	□
Q16 I thought the service was complicated.	□	□	□	□	□	□	□
Q17 I felt the service was easy to use.	□	□	□	□	□	□	□
Q18 I found the service frustrating to use.	□	□	□	□	□	□	□
Q19 I had to concentrate hard to use the service.	□	□	□	□	□	□	□
Q20 I did not feel in control when using the service.	□	□	□	□	□	□	□

Appendix 1.5. – Proposal Likert (PL) questionnaire

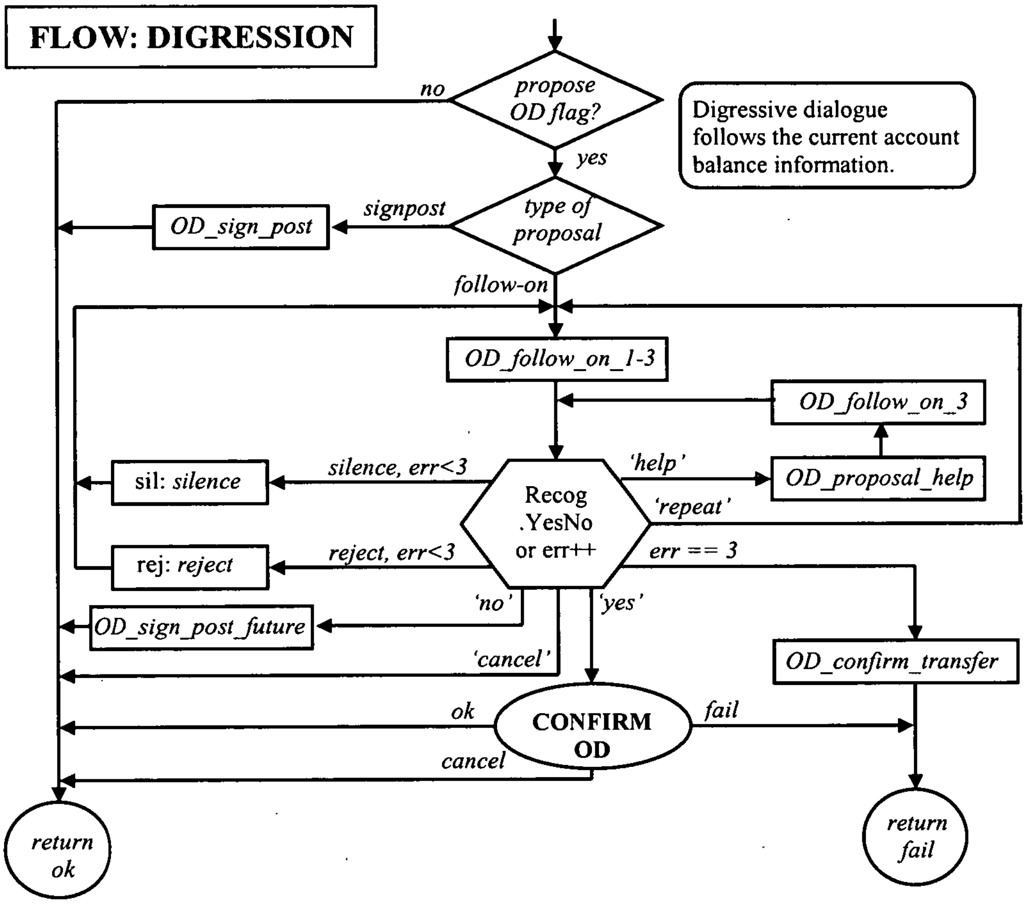
Thinking about the service you have just used. For each statement below, tick the box which best expresses your opinion on that statement.

	Strongly Agree	Agree	Slightly Agree	Neither agree nor disagree	Slightly Disagree	Disagree	Strongly Disagree
Q1 The overdraft proposal was annoying.	□	□	□	□	□	□	□
Q2 I now know how to use the automated service to apply for an overdraft.	□	□	□	□	□	□	□
Q3 The overdraft proposal information was helpful.	□	□	□	□	□	□	□
Q4 The overdraft proposal was too long.	□	□	□	□	□	□	□
Q5 I would trust the automated service to give me appropriate overdraft information.	□	□	□	□	□	□	□
Q6 I'd prefer an overdraft proposal to be made by a human agent rather than the automated service.	□	□	□	□	□	□	□
Q7 The overdraft proposal is an efficient method for giving product information.	□	□	□	□	□	□	□
Q8 I found the overdraft proposal intrusive.	□	□	□	□	□	□	□
Q9 I wouldn't rely solely on the automated service when seeking an overdraft.	□	□	□	□	□	□	□
Q10 The overdraft proposal was polite.	□	□	□	□	□	□	□
Q11 The overdraft proposal interrupted the call too much.	□	□	□	□	□	□	□
Q12 The overdraft proposal was easy to understand.	□	□	□	□	□	□	□
Q13 It is appropriate to have overdraft proposals in this kind of automated service.	□	□	□	□	□	□	□
Q14 The overdraft proposal distracted me from what I was trying to do.	□	□	□	□	□	□	□
Q15 If I needed an overdraft, I would be happy to apply through the automated service.	□	□	□	□	□	□	□
Q16 The overdraft proposal was irrelevant to me.	□	□	□	□	□	□	□

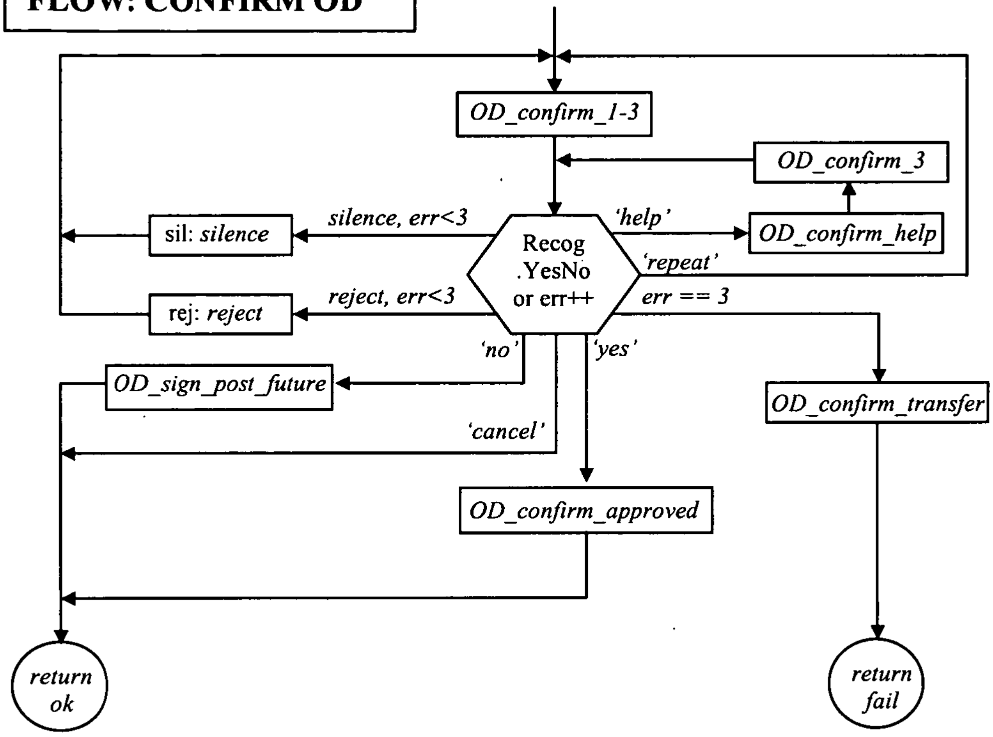
Appendix 2

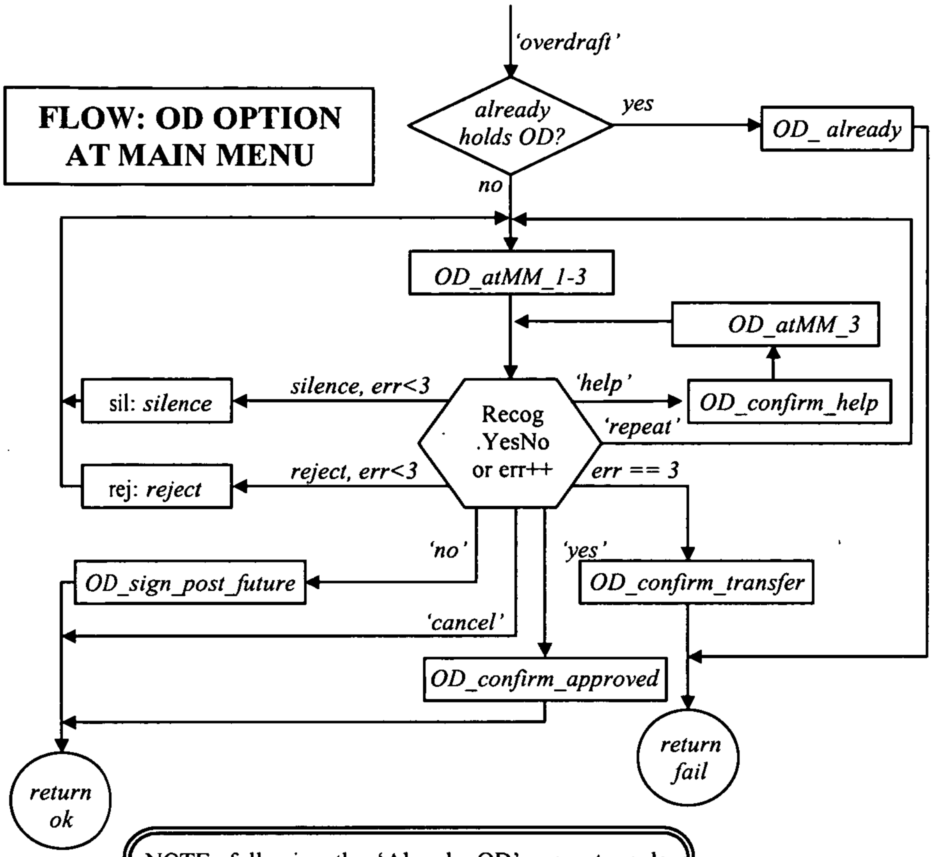
Appendix 2.1. – Experiment 1: Dialogue flow-charts for the system-initiated digressions

This section details the flow-charts and prompts involved in the 'Signpost' and 'Follow-on' digressive dialogue strategies.



FLOW: CONFIRM OD





FLOW: OD OPTION AT MAIN MENU

NOTE: following the 'Already_OD' prompt, replac 'S310-init' with 'OD_already_transfer' and 'S310 help' with 'OD_already_transfer_help'. For *yes* replie - transfer to advisor, for *no* replies - return to the Mai Menu stage.

Prompt	***Script***
OD_sign_post	You might like to know that you can have an overdraft of <OD amount> on this account. If you are interested, please say 'overdraft' at the main menu.
OD_follow_on_1	You might like to know that you can have an overdraft of <OD amount> on this account. Would you like this overdraft now?
OD_follow_on_2	*- use recordings for OD_atMM_2 -*
OD_follow_on_3	*- use recordings for OD_atMM_3 -*
OD_sign_post_future	If you would like to apply for an overdraft in the future, just say 'overdraft' at the main menu.
OD_confirm_transfer	I'm sorry, I'm having difficulty understanding if you would like an overdraft.
OD_proposal_help	At this point I need you to confirm if you would like the overdraft. *(followed by OD_proposal_3)*
OD_confirm_1	To confirm, you would like an overdraft of <OD amount>, is that correct?
OD_confirm_2	You would like an overdraft of <OD amount>, answering yes or no, is that correct?
OD_confirm_3	You can either say yes or press 1 on your telephone keypad, or say no or press 9. You would like an overdraft of <OD amount>, is that correct ?
OD_confirm_help	At this point I need you to confirm if you would like the overdraft. *(followed by OD_confirm_3)*
OD_confirm_approved	Thank you. Your <account name> account now has an overdraft limit of <OD amount>. You'll receive written confirmation within the next few days.
OD_already	You already have an overdraft limit of <OD amount> on your <account name> account. *(followed by OD_already_transfer)*
OD_atMM_1	You can have an overdraft of <OD amount> on your <account name> account. Would you like this overdraft now?
OD_atMM_2	You can have an overdraft of <OD amount> on your <account name> account. Would you like this overdraft now?
OD_atMM_3	You can either say yes or press 1 on your telephone keypad or say no or press 9. Would you like an overdraft of <OD amount> on your <account name> account?
OD_already_transfer	(follows Already_OD) We can transfer you to an Agent who can help you with your overdraft enquiry. Would you like to be transferred?
OD_already_transfer_help	At this point I need you to confirm if you would like to be transferred to an Agent who can help you with your overdraft enquiry. Either say yes or press 1 on your telephone keypad, or say no or press 9. Would you like to be transferred?

Appendix 2.2. – Experiment 1: Task sheet

Find out how much money is in your current account.

Find out if a cheque for £50 has been paid out of your current account.

The balance of the current account is: □ £______

□ Don’t know

Has the cheque been paid out of your current account?

□ Yes

□ No

□ Not sure

Appendix 2.3. – Experiment 1: Demographic questionnaire

Demographic/Technographic Questionnaire
(to be completed by researcher)

1. **Age:** □18-35 □ 36-49 □ 50+ **Gender:** □ male □ female

   Occupation:..
   (if student – parent's occupation)

2. **Accent** *(circle)*:

English	*north-west*	*north-east*	*midlands*	*south-west*	*south-east*
RP					
Scottish	*lowlands-east*	*lowlands-west*		*highlands*	
Welsh					
Irish	*north*	*south*			
Foreign					

additional comments:..

3. **How do you currently do most of your banking (tick all that apply)?**
   □ Visiting local branch □ PC/Internet banking □ ATM
   □ Phoning local branch □ Other:...
   □ Telephone banking (automated or talking to human)

4. **Have you ever used an automated telephone banking service before?**
   □ Yes □ No

   If yes, ask questions 5-8,

5. **Which automated telephone banking service do you use?** ....

6. **How often do you use automated telephone banking?**
   □ Once a day □ Once a week □ Once a month □ Never
   □ 2+ times a week □ 2+ times a month □ Less than once a month

7. **Do you use speech or touch-tone (the buttons on your telephone keypad)?**
   □ Speech
   □ Touch-tone

8. **What type of banking transactions do you do using automated telephone banking?**
   □ Balance Enquiry □ Recent transactions
   □ Pay Bills □ Transfer between accounts
   □ Other: ....

9. **Have you ever used any other automated telephone services before?**
   □ Yes - Which? ....
   □ No

Appendix 2.4. – Experiment 1: Exit questionnaire

Exit Questionnaire

(to be completed by researcher)

1. **Did you notice anything different with the last phone call to the service?**

   □ Yes □ No *If yes,* **what was different?**..

   ....

(if the participant answered 'no' or didn't mention 'overdraft' –> explain the overdraft offer)
(if the participant is still unaware about the overdraft offer -> continue with question 12)

2. **How did you feel about being offered an overdraft (your initial reactions)?**

   ....

3. **Did you take up the offer of the overdraft?**

   □ Yes □ No *If yes,* **how much is your overdraft and on which account?**..........

   ....

   If no, **how would you use the automated service to get an overdraft if you wanted one?**

   ....

4. **Why did you (or why did you *not*) decide to take up the overdraft offer?**

   ....

5. **Why do you think an overdraft was offered?**...

   ....

6. **Was it easy to understand the information in the overdraft message?**

   □ Yes □ No *If no,* **how could it be made clearer or improved?**.........................

   ....

7. **Would you have liked different/other information about the overdraft?**

   □ Yes □ No *If yes,* **what kind of information would you like, or how should the message be changed?** ....

8. **How do you imagine you would have reacted to the overdraft offer if this was a real service and you were using your own bank details (and why)?** ....

   ....

9. **If you were using this as a real service, would you prefer never to be offered an overdraft?**

   □ Yes □ No □ Other:..
   Any particular reasons why?...

10. **If this was a real service, would you want to be able to negotiate about the overdraft limit, e.g. perhaps choosing a lower/higher amount than was offered?**

□ Yes □ No □ Other:...
Comments?..

11. **If you were using this as a real service, would being offered an overdraft in this way discourage you from using the service in the future?**

□ Yes □ No *If yes,* **what is the strongest reason for this?**.................................

12. **I have a list of financial products that could be offered to you in this way through the service you just tried. I will read these products out to you and for each of them I would like you to answer whether you "would like"***(1)* **this product to be offered to you, "wouldn't mind"***(2)* **the product offer or if you "would dislike"***(3)* **the product offer?**

Personal loans:__
Credit cards:__
Overdrafts:__

Home insurance:__
Car insurance:__
Travel insurance:__

Accounts with high interest rates:__
Investments (e.g. ISA, bonds):__
Long term savings accounts:__

Would you like information in the service about a product that I haven't mentioned?..

13. **If this was a real service, which of the following three options would you prefer?**

□ A service which offers you products occasionally – like the one you just tried
□ A service providing information about financial products – only when you decide to select this
□ A service without any information about financial products at all

Comments (if any):..

14. **How do you prefer to find out about financial products (like loans or savings) in general?**

□ Information through the post
□ Information from Internet
□ Go to the branch
□ Receiving a phone call from your bank

□ Making a phone call to your branch
□ Information in the newspaper
□ Information on TV
□ Billboard advertising

□ Other:..

What is your preference and why?:..

15. **Are there any other comments about the service that you would like to add?**

....
....
....
....
....

Appendix 3

Appendix 3.1. – Experiment 2: Task sheet

Find out the **balance** of your current account.

The balance was: □ £________
□ Don't know

Order a **printed statement** for your savings account.

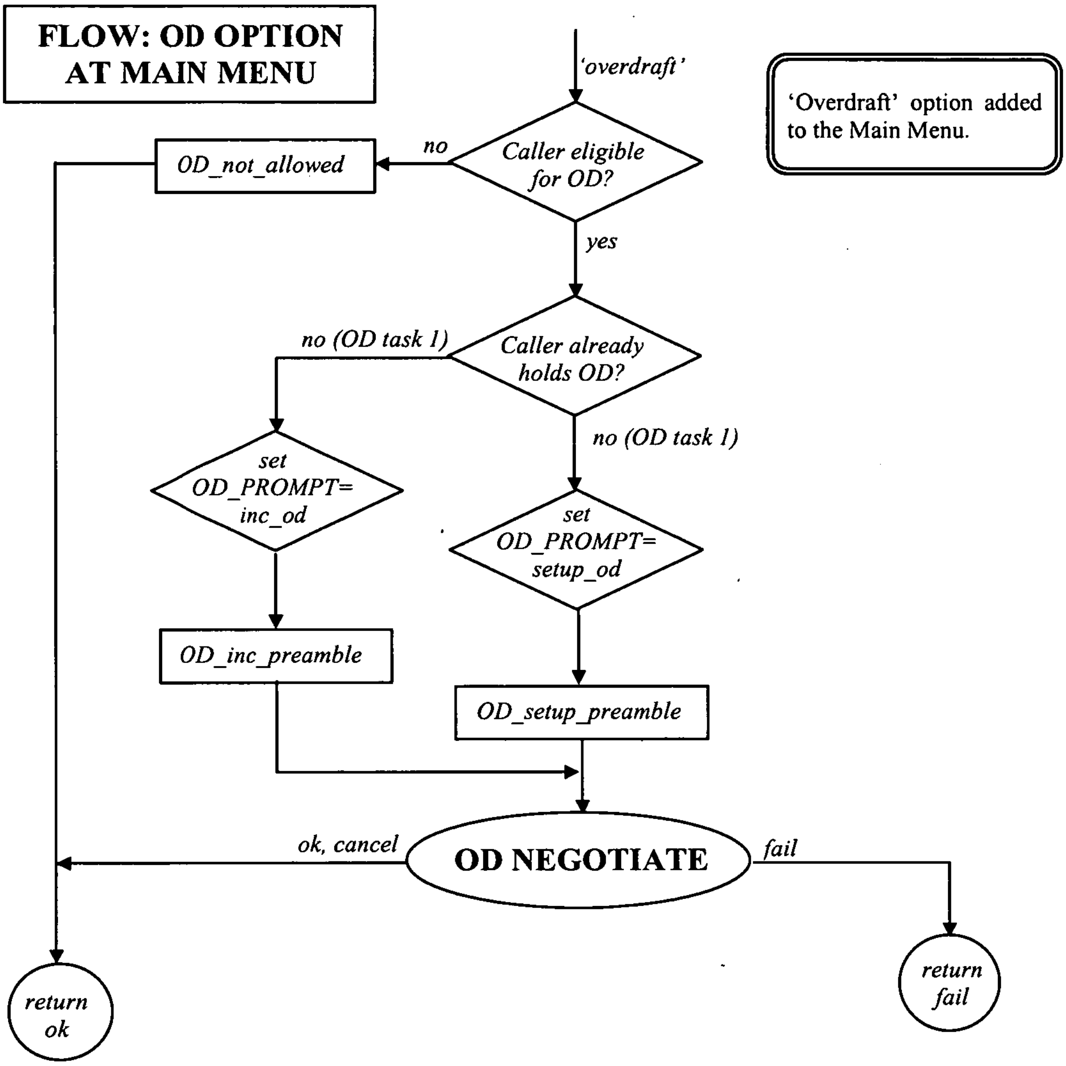
Apply for an **overdraft** on your current account.

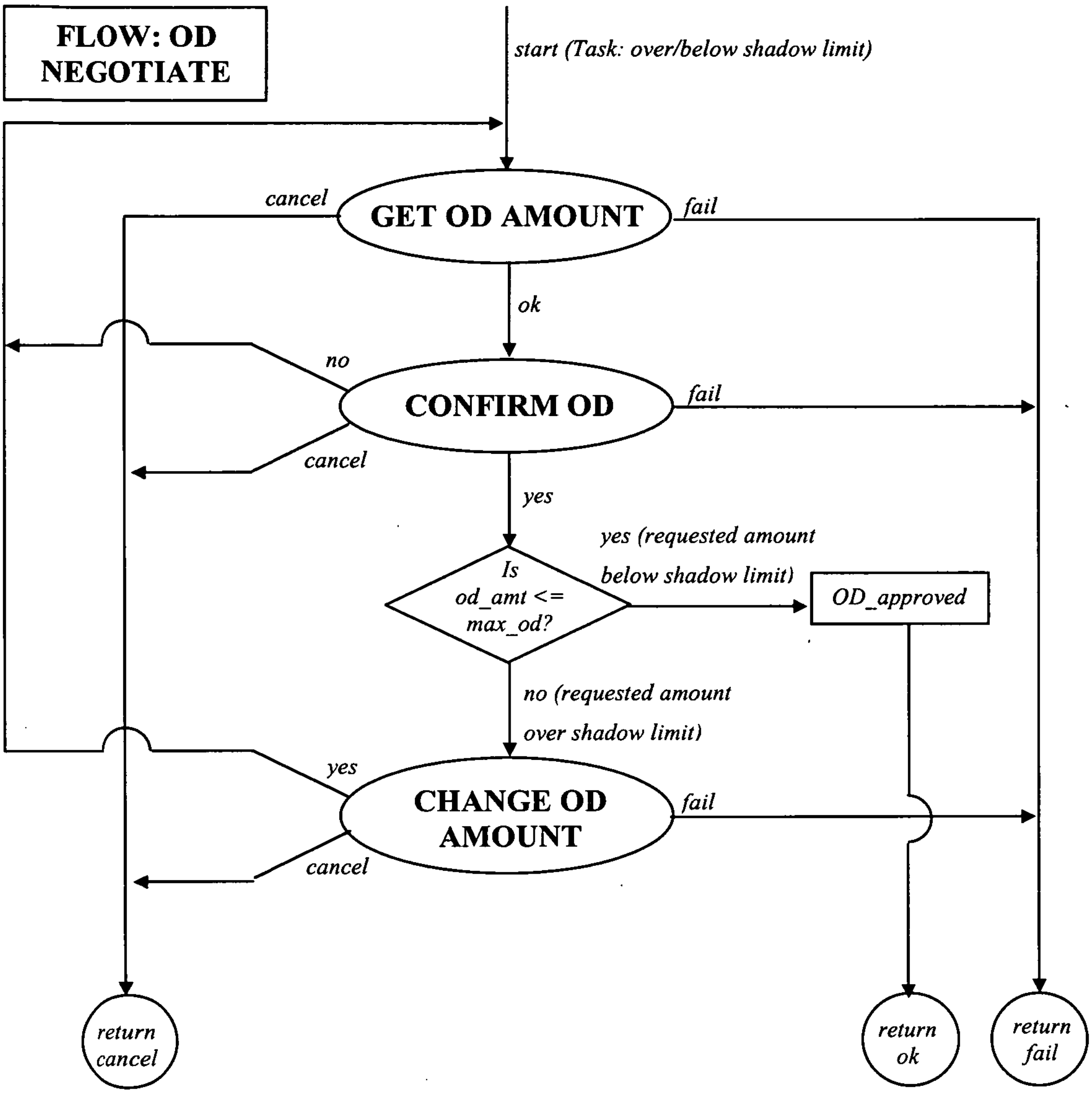
Did you get an overdraft limit on your account?
□ Yes
□ No
□ Don't know

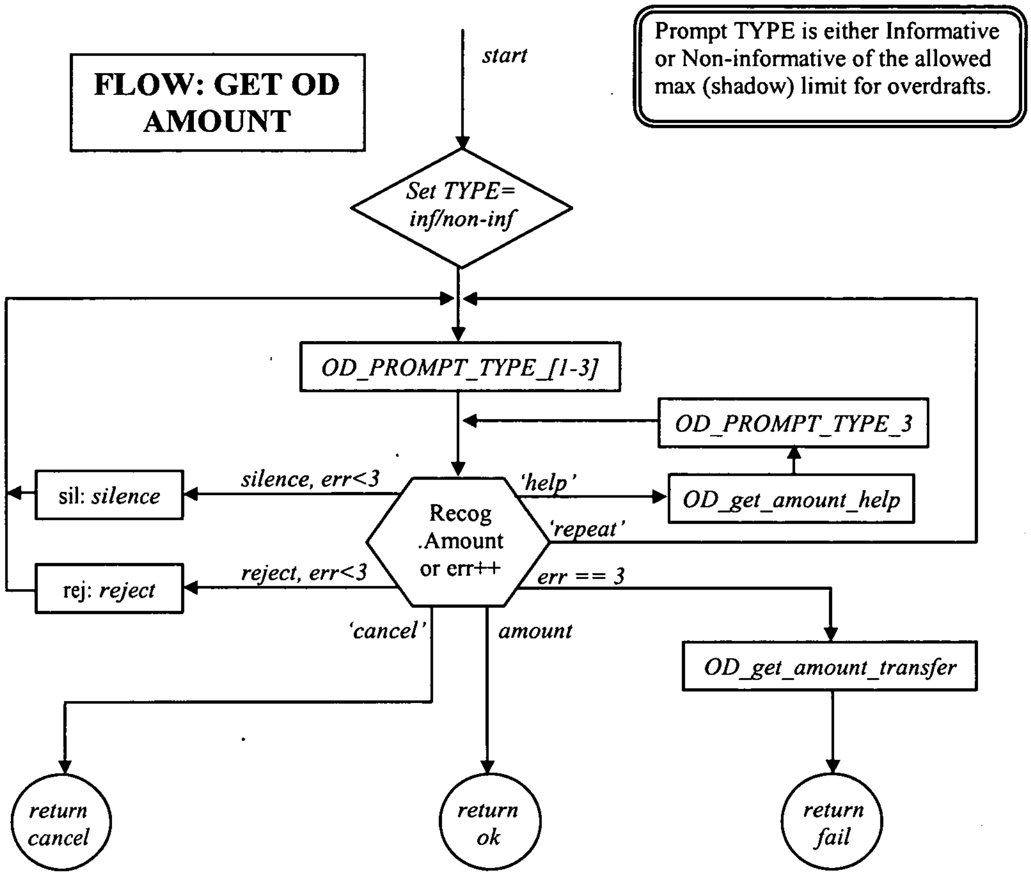
Appendix 4

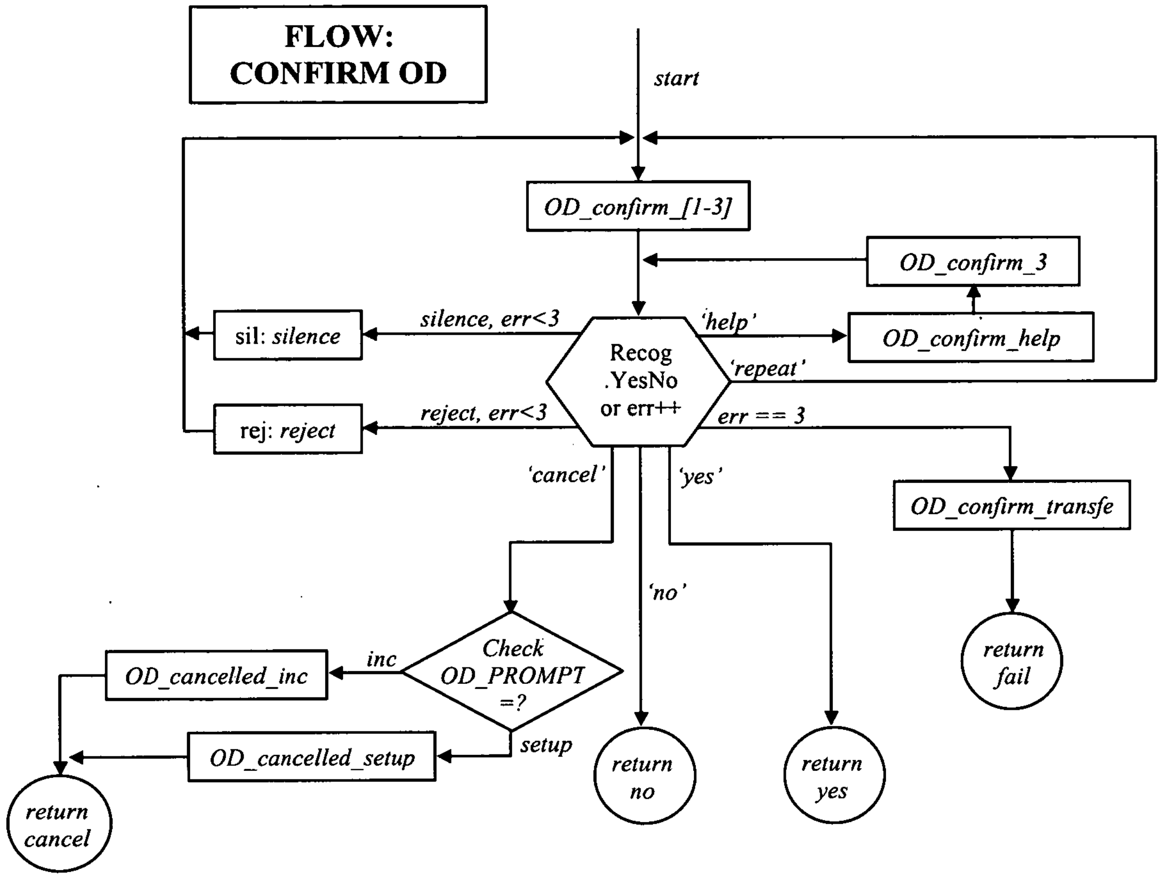
Appendix 4.1. – Experiment 3: Dialogue flow-charts

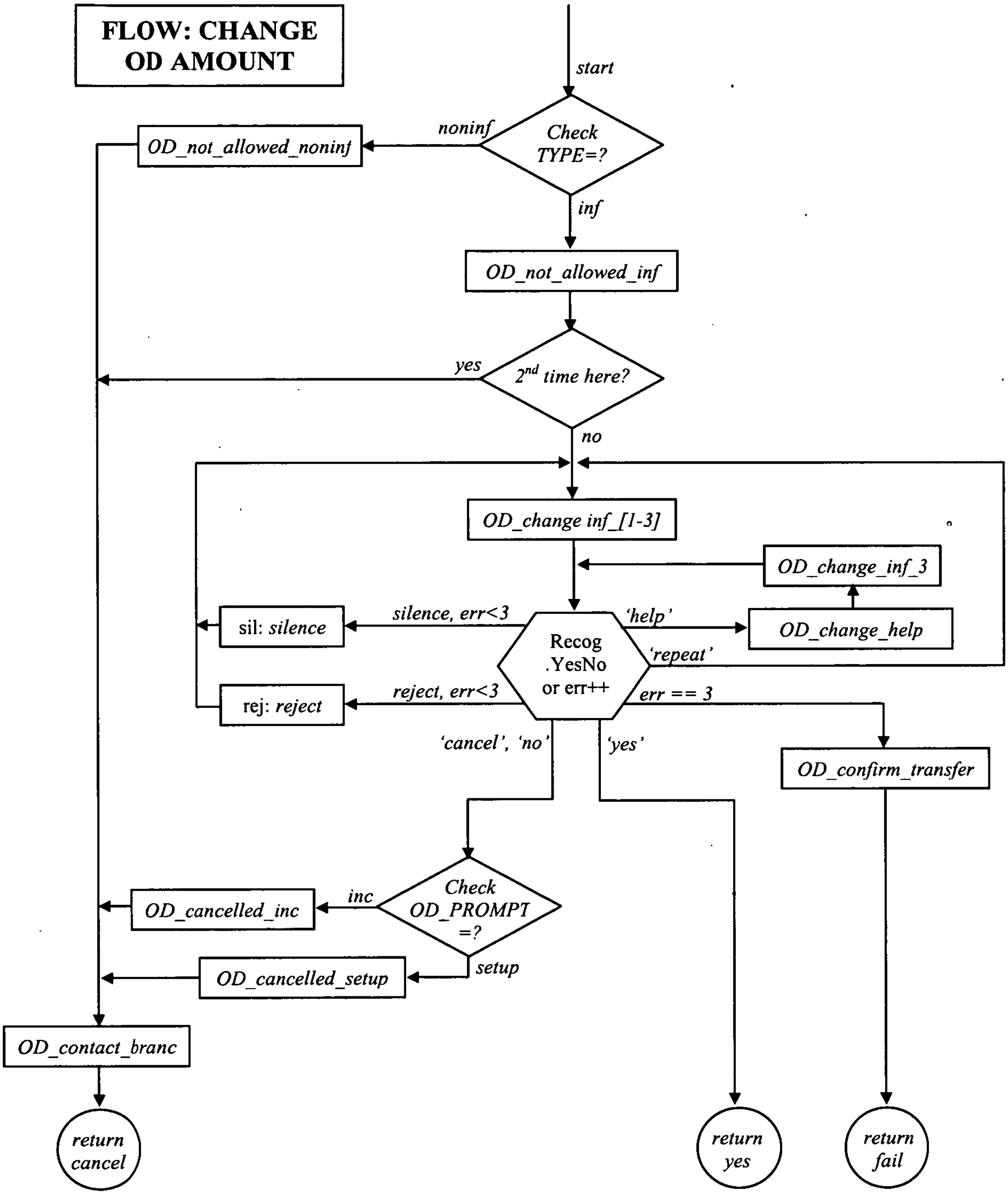
This section details the flow-charts and prompts involved in the 'Signpost' and 'Follow-on' digressive dialogue strategies and overdraft application tasks in Experiment 3.




FLOW: OD NEGOTIATE
start (Task: over/below shadow limit)
cancel
GET OD AMOUNT
fail
ok
no
CONFIRM OD
fail
cancel
yes
yes (requested amount
below shadow limit)
Is
od_amt <=
max_od?
OD_approved
no (requested amount
over shadow limit)
yes
CHANGE OD
AMOUNT
fail
cancel
return
cancel
return
ok
return
fail


FLOW: GET OD AMOUNT
start
Prompt TYPE is either Informative or Non-informative of the allowed max (shadow) limit for overdrafts.
Set TYPE= inf/non-inf
OD_PROMPT_TYPE_[1-3]
OD_PROMPT_TYPE_3
sil: silence
silence, err<3
'help'
OD_get_amount_help
Recog .Amount or err++
'repeat'
rej: reject
reject, err<3
err == 3
'cancel'
amount
OD_get_amount_transfer
return cancel
return ok
return fail


FLOW:
CONFIRM OD
start
OD_confirm_[1-3]
OD_confirm_3
silence, err<3
'help'
sil: silence
Recog
.YesNo
or err++
OD_confirm_help
'repeat'
reject, err<3
err == 3
rej: reject
'cancel'
'yes'
OD_confirm_transfe
'no'
Check
OD_PROMPT
=?
inc
OD_cancelled_inc
return
fail
setup
OD_cancelled_setup
return
no
return
yes
return
cancel


FLOW: CHANGE OD AMOUNT
start
Check TYPE=?
noninf
OD_not_allowed_noninf
inf
OD_not_allowed_inf
yes
2nd time here?
no
OD_change inf_[1-3]
OD_change_inf_3
'help'
OD_change_help
silence, err<3
sil: silence
Recog .YesNo or err++
'repeat'
reject, err<3
rej: reject
err == 3
'cancel', 'no'
'yes'
OD_confirm_transfer
Check OD_PROMPT =?
inc
OD_cancelled_inc
setup
OD_cancelled_setup
OD_contact_branc
return cancel
return yes
return fail

Prompt	*Script*
OD_setup_preamble	"We can set up an overdraft facility on your current account today. You can use as much of your overdraft facility as you need when you need it, and there are no charges if your account is overdraft by less than ten pounds. If you take up the overdraft facility we'll send you a confirmation letter with details of terms and conditions".
OD_inc_preamble	"The present overdraft limit on your current account is ... pounds."
OD_sign_post	You might like to know that you can have an overdraft of <OD amount> on this account. If you are interested, please say 'overdraft' at the main menu.
OD_follow_on_1	You might like to know that you can have an overdraft of <OD amount> on this account. Would you like this overdraft now?
OD_follow_on_2	*- use recordings for OD_atMM_2 -*
OD_follow_on_3	*- use recordings for OD_atMM_3 -*
OD_sign_post_future	If you would like to apply for an overdraft in the future, just say 'overdraft' at the main menu.
OD_confirm_transfer	I'm sorry, I'm having difficulty understanding if you would like an overdraft.
OD_proposal_help	At this point I need you to confirm if you would like the overdraft. *(followed by OD_proposal_3)*
setup_od_inf_1	You can have an overdraft limit of up to <max OD amount>. How much would you like?
setup_od_inf_2	You can have an overdraft limit of up to <max OD amount> on your <account name>, how much would you like your overdraft limit to be?
setup_od_inf_3	The maximum overdraft limit available on your <account name> through this service is <max OD amount>. Please say how much you would like your overdraft limit to be.
setup_od_noninf_1	How much would you like your overdraft limit to be?
setup_od_noninf_2	You-can have an overdraft facility on your <account name>. How much would you like your overdraft limit to be?
setup_od_noninf_3	You can have an overdraft facility on your <account name>. Please say how much you would like your overdraft limit to be.
inc_od_inf_1	You can have an overdraft limit of up to <max OD amount>. How much would you like?
inc_od_inf_2	You can have an overdraft limit of up to <max OD amount> on your <account name>, how much would you like your overdraft limit to be?
inc_od_inf_3	The maximum overdraft limit available on your <account name> through this service is <max OD amount>. Please say how much you would like your overdraft limit to be.

inc_od_noninf_1	How much would you like your new overdraft limit to be?
inc_od_noninf_2	You can arrange a new overdraft limit on your <account name>.How much would you like your overdraft limit to be?
inc_od_noninf_3	You can arrange a new overdraft limit on your <account name>.Please say how much you would like your overdraft limit to be.
OD_get_amount_help	At this point I need you to say an amount in whole pounds. *(followed by OD_PROMPT_TYPE_3)*
OD_get_amount_transfer	I'm sorry. I'm having difficulty understanding how much you would like your overdraft limit to be.
OD_confirm_1	To confirm, you would like an overdraft of <OD amount>, is that correct?
OD_confirm_2	You would like an overdraft of <OD amount>, is that correct?
OD_confirm_3	You can either say yes or press 1 on your telephone keypad, or say no or press 9. Would you like an overdraft of <OD amount>?
OD_confirm_help	At this point I need you to confirm that I heard the overdraft amount correctly. *(followed by OD_PROMPT_TYPE_3)*
OD_approved	Thank you. Your <account name> account now has an overdraft limit of <OD amount>. You'll receive written confirmation within the next few days.
OD_cancelled_setup	Your overdraft request has been cancelled.
OD_cancelled_inc	Your overdraft limit remains unchanged at <OD limit>.
OD_confirm_transfer	I'm sorry. I'm having difficulty understanding if the amount is correct.
OD_change_inf_1	Would you like to arrange an overdraft within that limit?
OD_change_inf_2	Would you like to arrange an overdraft within the limit of <max OD amount>?
OD_change_inf_3	You can either say yes or press 1 on your telephone keypad or say no or press 9. Would you like to arrange an overdraft within the limit of <max OD amount>?
OD_change_help	At this point I need to know whether you would like to arrange an overdraft limit lower than the one you previously requested.
OD_change_transfer	I'm sorry. I'm having difficulty understanding if you would like to arrange a lower overdraft limit.

Appendix 5

Appendix 5.1. – Experiment 4: SID questionnaire

Thinking about the <u>proposal in the service you have just used</u>. For each statement below, tick the box which best expresses your opinion on that statement.

	Strongly Agree	Agree	Slightly Agree	Neither agree nor disagree	Slightly Disagree	Disagree	Strongly Disagree
Q1 The proposal was annoying.	□	□	□	□	□	□	□
Q2 The wording of the proposal fitted in well with the rest of the service.	□	□	□	□	□	□	□
Q3 The style of the proposal was too formal.	□	□	□	□	□	□	□
Q4 The proposal was too long.	□	□	□	□	□	□	□
Q5 I would trust the service to give me appropriate information about savings accounts.	□	□	□	□	□	□	□
Q6 The proposal made me feel I was being manipulated.	□	□	□	□	□	□	□
Q7 The proposal was an efficient way of giving information about the Online Saver account.	□	□	□	□	□	□	□
Q8 I found the proposal intrusive.	□	□	□	□	□	□	□
Q9 I would want more information before opening an Online Saver account.	□	□	□	□	□	□	□
Q10 The proposal was polite.	□	□	□	□	□	□	□
Q11 The proposal interrupted the call too much.	□	□	□	□	□	□	□
Q12 The proposal was easy to understand.	□	□	□	□	□	□	□

Please turn over...

	Strongly Agree	Agree	Slightly Agree	Neither agree nor disagree	Slightly Disagree	Disagree	Strongly Disagree
Q13 The proposal was appropriate for this service.	□	□	□	□	□	□	□
Q14 The proposal distracted me from what I was trying to do.	□	□	□	□	□	□	□
Q15 The proposal was friendly.	□	□	□	□	□	□	□
Q16 The proposal contained only relevant information.	□	□	□	□	□	□	□
Q17 The proposal expressed care for my individual needs.	□	□	□	□	□	□	□
Q18 The proposal should give more information about the Online Saver account.	□	□	□	□	□	□	□
Q19 The proposal information was helpful.	□	□	□	□	□	□	□
Q20 The proposal was very long-winded.	□	□	□	□	□	□	□
Q21 The proposal took my interests into account.	□	□	□	□	□	□	□
Q22 If I wanted an Online Saver account, I would be happy to open it through the automated service.	□	□	□	□	□	□	□
Q23 I found the proposal patronising.	□	□	□	□	□	□	□
Q24 The way the proposal was expressed was too apologetic.	□	□	□	□	□	□	□

Appendix 5.2. – Experiment 4: FRS questionnaire

Thinking about the proposal I've just heard, it was...

polite ___|___|___|___|___|___|___ impolite

informal ___|___|___|___|___|___|___ formal

to the point ___|___|___|___|___|___|___ long-winded

forthright ___|___|___|___|___|___|___ diplomatic

insincere ___|___|___|___|___|___|___ sincere

patronising ___|___|___|___|___|___|___ respectful

personalised ___|___|___|___|___|___|___ impersonal

apologetic ___|___|___|___|___|___|___ unapologetic

I associate the choice of wording in the proposal with someone who is...

tactful ___|___|___|___|___|___|___ tactless

timid ___|___|___|___|___|___|___ self-confident

unsociable ___|___|___|___|___|___|___ sociable

reliable ___|___|___|___|___|___|___ unreliable

caring ___|___|___|___|___|___|___ uncaring

unprofessio ___|___|___|___|___|___|___ professional

Publications

Journal: Behaviour and Information Technology (In press)

Hidden menu options in automated human-computer telephone dialogues: dissonance in the user's mental model

JENNY WILKIE*[†], FERGUS MCINNES[†], MERVYN A. JACK[†], and PETER LITTLEWOOD[‡]

[†]Centre for Communication Interface Research, UK
University of Edinburgh
AGB Building, King's Buildings, Edinburgh, UK, EH9 3JL
tel: +44 131 650 2801
fax: +44 131 650 2784
{jenny/fmi/maj}@ccir.ed.ac.uk

[‡]Lloyds TSB Telephony, UK
Canons Way, Bristol, UK, BS99 7LB
peter.littlewood@lloydstsb.co.uk

Abstract

This paper explores the consequences of adopting an alternative strategy to that of explicitly listing all options within the main menu of a speech-driven automated telephone banking service. An existing service was augmented with an overdraft request dialogue, accessible at its main menu, which could be triggered using the keyword "overdraft". However, the keyword was not explicitly mentioned as an option in the main menu. Instead, system-initiated proposals for an overdraft were introduced into the call flow notifying callers that they could apply for an overdraft by saying "overdraft" at the main menu. An experiment with 114 participants was carried out to investigate the effectiveness of this strategy as a way of offering new services without increasing the length of the main menu. Results showed that a significant proportion of participants (37%) did not succeed in completing an overdraft request. The reasons for this failure are discussed.

Keywords: Human-computer dialogues; menus; speech recognition

1. Introduction

Speech recognition is increasingly used in the mass-market domain of self-service telephone applications. Compared to their touch-tone counterparts, speech enabled telephone applications offer users a more natural and flexible way of interacting with a computer-based system. However, whilst the capability of speech recognition technology is continuously improving, the need still exists for the system messages (or prompts) to be designed to constrain the range of user inputs to those that match the capabilities of the speech recognition grammar (Karis and Dobroth 1991, Bernsen et al. 1996, Tomko 2004). Much of the dialogue design for mass-market automated telephone services is centred around making the interaction fit the technology at hand, relying on explicit (or implicit) instructions and error recovery strategies in order to guide users as to what to say, how to say it and when to speak.

The system functionality remains hidden to the user in auditory-only interfaces and a central concern in the design of such applications is therefore how to let callers know about the range of available options (Yankelovich 1996). Several studies have been conducted to explore how contrasting menu prompt strategies may affect user satisfaction and task completion in applications such as: call routing (Sheeder and Balogh 2003, Williams et al. 2003a, b), telephone directory assistance (Vanhoucke et al. 2001), telephone banking (McInnes et al. 1999), e-mail (Walker et al. 1998) and newspaper subscriptions (Dialogues 2000 Report, 1997). These studies have centred around three main strategies: 1) 'open' prompts[1] (e.g. "How may I help you?") inviting the user to say any utterance; 2) 'closed' prompts[2] where the user makes a selection from an up-front list of options; or 3) on-request prompts where the user is not presented with options unless specifically requesting this information, for example by using commands such as "help" or "hear list". Combinations of these three strategies have also featured, such as presenting the caller with an open prompt initially and then, if the speech recogniser detects no response (silence), playing the list of menu options or giving further instructions (see Sheeder and Balogh 2003).

There is no universally applicable strategy to draw on when designing prompts for voice-driven telephone applications. The ideal strategy will depend on the skills of the intended user group, the application domain, the frequency of use and the complexity of the underlying data structure (Vanhoucke et al. 2001). The majority of today's commercially deployed mass-market automated applications are aimed at servicing the general public and must therefore provide for callers with diverse levels of experience and knowledge. In these types of self-service applications, the use of menus in the form of explicit list selections is the most popular and frequently employed method for informing users (especially novice users) about the range of services available to them. Although listening generally requires less perceptual and cognitive effort than reading (Preece et al. 2002: 87), information presented through the auditory-only interface is serial, transient and paced, which puts a strain on users'

[1] The degrees of 'openness' may vary. For example, the prompt "How may I help you?" can be made more specific by suggesting to the user the type of utterances which are expected "Which *service* do you require?" (see McInnes et al. 1999, Cohen et al. 2004).

[2] Particularly pronounced forms of closed prompts are featured in dialogues where user inputs are restricted to responding with "yes" or "no" to each menu option or, as in the 'skip and scan' strategy, by using a predefined command, e.g. "next", "previous" or "select" (see Hornstein 1994, Dialogues 2000 Report 1997, Goldstein et al. 1999).

cognitive and perceptual resources. This has an impact on the length of system prompts and limits the number of options that can usefully be presented in menus[3].

Touch-tone key mappings for menu options are also often provided as an alternative input strategy alongside voice, for example, when the user may prefer to push telephone buttons (e.g. giving a sense of privacy when entering bank account information) or when the human-computer interaction needs supporting in order to avoid a breakdown (e.g. as fall-back after repeated mis-recognitions, or in noisy environments). The inclusion of touch-tone key mappings, coupled with the fact that system prompts frequently consist of pre-recorded human speech, lead to rather rigid application structures where changes once the service has been implemented and launched are impractical and can be costly. A dilemma facing designers of such menu-based applications is where to place new or less frequently requested service options within the call-flow and how to incorporate these with the existing interface design. Voice recognition design guidelines described in the recent literature (Weinschenk and Barker 2000, Kotelly 2003, Cohen et al. 2004) offer broad coverage for how to implement menus in mass-market automated telephone services. However, they do not fully address issues surrounding the maintenance and future development of such menu-driven services.

There are a number of reasons for looking beyond the conventional menu-based dialogue design to explore alternative and more flexible means of offering users access to services through an automated application. For example, an enterprise may want to introduce new informational or transactional services that may normally not be considered in the initial application design, such as access to services which are infrequently requested, short-term offers or product promotions. Furthermore, some facilities within the application may only apply to certain callers under particular circumstances, and adding these to the listings of the core service options may render menu structures cluttered and complex.

One clear advantage with voice-based interfaces (as opposed to touch-tone) is that no key-mappings for service options need to be explicitly conveyed to the user through menus; instead of pressing a key the user can invoke a service option by stating the name of it, and often by using synonymous expressions. This allows for a menu option to remain active in the application but 'hidden' until the user requests it. Still, callers need to be made aware about the range of options that are available to them through the service. One solution is to introduce a short one-off system-initiated message at a relevant point within the dialogue structure to let the caller know that the (hidden) service option is available and how to select it. This system-initiated message may or may not be followed by a short dialogue (e.g. requesting a yes/no response) enabling the user to pursue or decline the offer immediately.

Previous research by the authors (Wilkie et al. 2002) identified two key dialogue engineering issues for the design of such system-initiated proposals in human-computer telephone dialogues: the *location* of the proposal in the application call-flow; and the dialogue *turn-taking strategy* employed for delivering the proposal information. In order to address these dialogue design issues, two experiments were devised in which system-initiated proposals were introduced in the call-flow of a

[3] For further information on cognitive load design consideration for voice interfaces, see Cohen et al. 2004: 119-131.

speech-driven automated telephone banking service (Wilkie et al. 2002). These proposals informed users that they could apply for an overdraft facility on their account through the automated service by saying "overdraft" at the menu of services, where the overdraft option was active but not explicitly listed. Participants' (n=114) attitudes towards the telephone banking service were measured before and after they were subjected to this additional overdraft proposal. In sum, the results from that research revealed that participants' attitudes towards the usability of the service remained unaffected by the delivery of the overdraft proposal.

Coupled with user acceptance and application usability, the future success of these system-initiated proposals will rely largely on the users' ability to successfully locate and select the hidden menu option within the automated service dialogue. This paper explores the participant performance data (task completion and navigational route through the dialogue) obtained in the proposal location experiment described in the preceding paragraph.

2. Method

2.1. Objectives

The results described here were obtained during an experiment in which system-initiated overdraft proposals were introduced at three different locations in an existing voice-driven automated telephone banking service (user attitudes towards these proposals are presented in Wilkie et al. 2002). The characteristic of the telephone banking service used in this research is that it is heavily reliant on menus and directed prompts for instructing users what to say. The analysis described here focuses on the users' ability to infer (once they have heard the system-initiated proposal) that they can say "overdraft" at the main menu of services --- even though this option is not explicitly listed.

2.2. Participants

Participants were recruited from the general public in Edinburgh. In total, 114 complete participant data sets were obtained and used in the statistical analysis. Previous exposure to an automated telephone banking service was not a prerequisite for taking part in the experiment, however, a significant proportion of the participant cohort[4] (45.5%) stated that they had used such a service for their personal banking, and of which eight (16%) said that the service could also recognise speech input. Five participants had used the target application employed in the experiment (application description is provided in the Apparatus section 2.3 below).

Participants received an honorarium payment of £20 for partaking in the research. Participant demographics are presented in table 1. There was a bias towards the younger age group in the sample, due to issues in the recruitment process.

[4] Four participants did not provide an answer to whether or not they had used an automated service.

	Age 18-35	Age 36-49	Age 50+	TOTAL
Female	44 (38.6%)	5 (6.1%)	15 (13.2%)	64 (56.1%)
Male	31 (27.2%)	7 (4.4%)	12 (10.5%)	50 (43.9%)
TOTAL	75 (65.8%)	12 (10.5%)	27 (23.7%)	114 (100%)

Table 1. Distribution of participants by age group and gender.

2.3. Apparatus

The automated telephone banking service used in this research was modelled on an existing real-world application which provides registered customers with access to personal account information (e.g. balance information or recent transactions) and enables them to perform a number of banking transactions (e.g. funds transfers or ordering account statements). The service accepts both speech and push-button caller input. There are no barge-in capabilities for voice inputs, which means that callers have to wait for the system prompt to finish playing before they can start speaking.

The banking dialogue was implemented using the commercially available Nuance (v7.0.3) speech recogniser software; it comes pre-packaged with language models and therefore does not require training data (i.e. speaker independent). For the purpose of the experiment, the Nuance Grammar Specification Language notation was employed to define word string representations of all permissible speech inputs (along with corresponding touch-button options) for each input stage in the dialogue, resulting in hand-crafted finite-state grammars against which caller responses are then compared to find a best match. The grammars also enable users to employ spoken natural language input by allowing for some extraneous speech (such as "Can I have ..., please."); enabling multiple pieces of information to be stated at once at the main menu (e.g. "I'd like the *balance* of my *current account* please."); and allowing callers to use synonymous phrases to express their requests (e.g. "balance" vs. "what's in my account?"). The grammar then returns the result of the recognition (i.e. success/failure, speech/touch-button input) and a 'semantic interpretation' of the caller's response, consisting of key information-bearing words required for the dialogue to progress to the next input/output stage.

The system setup had been used in a previous experiment (using the same banking dialogue and grammars) and had achieved a recognition accuracy of 94.3% on in-grammar utterances (with 4.1% rejected speech input and 1.5% misrecognitions). On out-of-grammar utterances, which comprised 9.6% of all speech input, 22.2% were misrecognised and 77.8% rejected.

Each dialogue stage (where input from the caller is processed) featured a generic three-level error recovery prompting strategy with incremental instructions which promoted speech; this meant that the first-level prompt instructions were brief, the second-level error recovery prompt gave the caller further details about the kind of input the system expected, and the final third-level error recovery prompt further revealed which push-button options to use. An error could be caused by either a silence or an invalid input rejected by the system, and callers were given three attempts at giving a recognisable input after which the call would be transferred to a human agent for further assistance. For the purpose of the experiment, however, the call transfer was replaced with a message informing participants to hang up and inform the researcher that the call had been transferred. The researcher then explained to the participant that their call would, in real life, be transferred to a human agent who would be able to help them with their enquiry.

The application dialogue is outlined in figure 1, showing the system prompts and user responses for the identification and verification stage, followed by a balance enquiry. A more detailed flow-chart of the identification dialogue (with associated error level prompts) is included in the Appendix. The prompts were recorded in a female Southern British English accent.

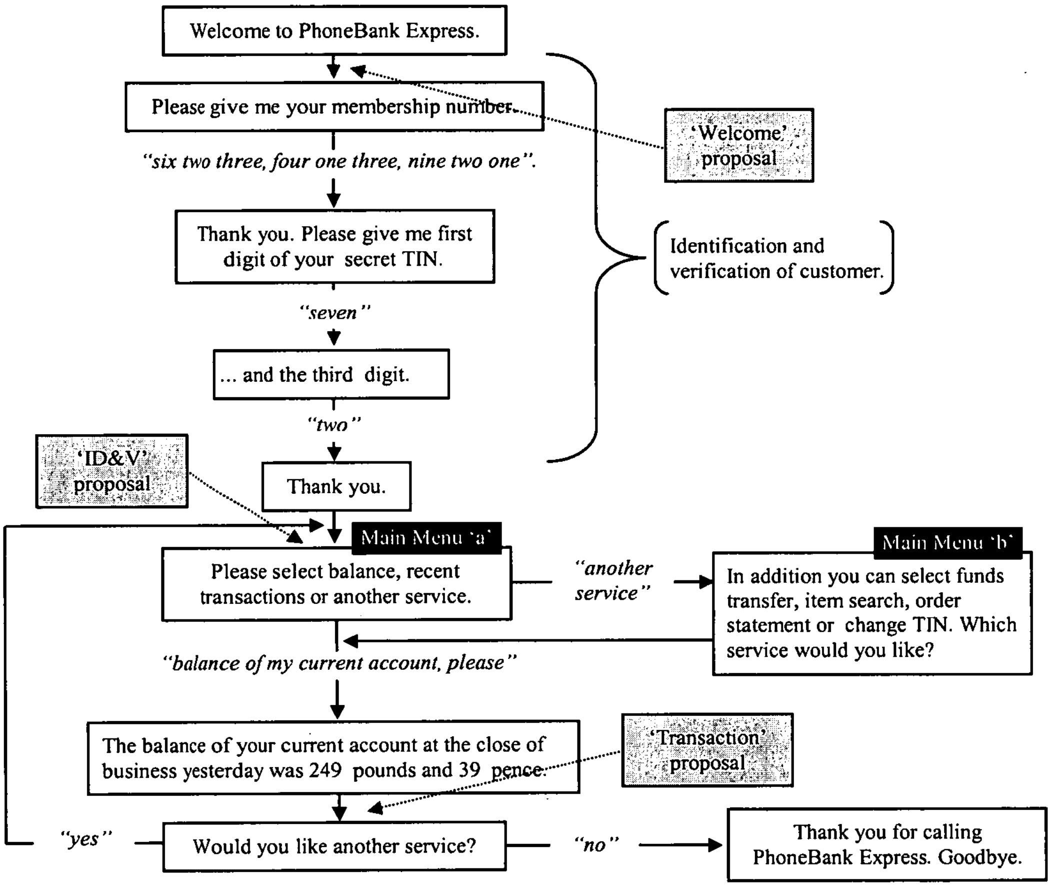

Welcome to PhoneBank Express.
Please give me your membership number.
"six two three, four one three, nine two one".
Thank you. Please give me first digit of your secret TIN.
"seven"
... and the third digit.
"two"
Thank you.
'Welcome' proposal
Identification and verification of customer.
'ID&V' proposal
Main Menu 'a'
Please select balance, recent transactions or another service.
"another service"
Main Menu 'b'
In addition you can select funds transfer, item search, order statement or change TIN. Which service would you like?
"balance of my current account, please"
The balance of your current account at the close of business yesterday was 249 pounds and 39 pence.
'Transaction' proposal
"yes"
Would you like another service?
"no"
Thank you for calling PhoneBank Express. Goodbye.

Figure 1. Flow-chart overview of the banking dialogue. System messages appear in boxes, and user responses are italicised. The location of the three system-initiated proposals are marked out using grey boxes. Abbreviations used in the flow are: TIN (Telephone Identification Number) and ID&V (Identification and Verification).

Service options were presented to callers at the main menu in the dialogue by the use of a two-tiered approach (these are labelled 'Main Menu a' and 'Main Menu b' in figure 1). The first half (Main Menu 'a') listed the most frequently requested service options; by saying "other services" the caller could bring up the second half of the listing (Main Menu 'b'). All service options were active for the caller to select at either half of the menu. This means that users could pre-empt, or volunteer, input information such as "order statement" already at the Main Menu 'a' stage (before the option had been explicitly listed). Throughout this paper, the term 'volunteered' will be used to refer to this particular dialogue behaviour. The overdraft option, which was considered a less frequently accessed service option, was active at the main menu (both the 'a' and 'b' stage) but was not explicitly listed. This means that the caller has to deduce that the option is available at the main menu rather than selecting it from the listing.

Once an enquiry has been completed within the service, the caller is prompted with *"Would you like another service?"*. At this stage the caller can do one of the following: respond "yes" to return to the Main Menu 'a' listing; immediately respond which service option that they require (thus bypassing the menu listings altogether); or respond with "no" to exit the service. The human-computer dialogue for a typical balance and order statement task for a customer with two accounts (a current and a savings account) are presented in figure 2.

Balance request	**Statement order**
C: Please select balance, recent transactions or another service. U: Balance. C: Is that for your current account? U: Yes. C: The balance of your current account is 286 pounds and 54 pence.	C: Please select balance, recent transactions or another service. *U: Another service. *C: In addition you can select funds transfer, item search, order statement or change TIN. Which service would you like? U: Order statement. C: Is that for your current account? U: No. C: Thank you. A statement for your savings account has been ordered.

Figure 2. Task examples of the prompt-response interaction in the banking service between computer (C) and user (U). Turn-taking stages marked with '*' can be bypassed.

An overdraft proposal was introduced in one of three different locations in the banking dialogue. The main characteristics of the proposals are a short 'signpost' style prompt, embedded within the normal dialogue call flow, informing customers about the availability of the overdraft service and how to obtain the overdraft within the automated dialogue --- in this case, by saying "overdraft" at the menu of services. The wordings of the proposals were optimised based on their location in the dialogue. The first of these, referred to here as the 'Welcome' proposal, followed immediately after the initial "welcome to PhoneBank *Express*" prompt. The Welcome proposal was worded in such a way as to be general in nature and applicable to all callers: *"We've added a new overdraft facility to this service. To find out more, just say overdraft at the menu of services"*. The second proposal variant was located immediately after a caller had been successfully identified by a valid membership number and verified successfully using their secret TIN (Telephone Identification Number). This version is referred to here as the 'ID&V' (identification and verification) proposal. Since the caller had been identified as this point in the dialogue, the proposal was made customer oriented with reference to specific amounts and accounts: *"You might like to know that you can have an overdraft of 400 pounds on your current account. To find out more, just say overdraft at the menu of services"*. Finally, the third 'Transaction' proposal was a nested prompt that followed the dialogue for the balance request, immediately after the amount had been played: *"You might like to know that you can have an overdraft of 400 pounds on this account. To find out more, just say overdraft at the menu of services"*.

2.4. Setup

The automated telephone banking application was installed on a PC with a PII 400 MHz dual processor and 256 MB of RAM, running Windows NT 4.0. A Dialogic D/300SC-E1 board (a 30-port DSP-based voice board with onboard digital E-1 ISDN telephone interface) was installed on the PC to run the telephony.

Participants were seated at a desk throughout the experiment and operated the automated banking service using a standard landline telephone with touch-tone buttons on the base of the telephone. Performance data were collected automatically throughout the dialogue. System log files were used to record, for each phone call, the system output (dialogue stage and error level) and the user input (speech/touch-tone and recognition results). Additionally, sound files were used to record user utterances.

2.5. Procedure

On arrival, participants were greeted by a researcher and informed that they would make some phone calls to an automated telephone banking service. No details of the main purpose of the current research (i.e. about the overdraft proposals, mental models of the automated service and the ability to navigate through the dialogue) were disclosed. Prior to making their first phone call to the service, participants were told that they could both speak their commands and use the buttons on the telephone keypad; however, no further instructions on how to use the service (i.e. which buttons to press or which words to use) were given. Each participant was then presented with a sheet of paper containing their (fictitious) persona details to be used throughout the experiment: a membership number, a TIN, the telephone number for the service and details of two accounts: a current account and a savings account. For ethical and data protection reasons, no personal data were used at any point in the experiment.

The participants' tasks were to make telephone calls (four in total) to the automated service to find and take a written note of the balance of 'their' current account and then, in the same phone call, to order a statement for 'their' savings account. These same two tasks were repeated for each of the four phone calls; between calls participants were instructed to imagine that "a few days have gone by" before making the next phone call. To encourage participants' involvement in the tasks the balance information changed between phone calls to the service. The first two phone calls allowed participants to become familiar with the service functionality and their persona details. In the third phone call, participants experienced an overdraft proposal (participants had not been forewarned about the proposal delivery). In the fourth and final call to the service, participants were asked, in addition to the balance and statement tasks, to apply also for an overdraft on the current account and to write down the result (whether an overdraft limit on their account had been obtained or not). The experimental procedure was kept consistent and identical for all participants, with exception of the location of the overdraft proposal in the third phone call. Participants were allowed up to three attempts to complete each phone call. A phone call was considered completed once the participant had successfully given a membership number and TIN and arrived at the menu of services in the dialogue (Main Menu 'a' in figure 1).

2.6. Design

There were three independent variables in this research: age group (three levels), gender, previous exposure to automated telephone banking (two levels) and location of overdraft proposal (three levels). A between-group experimental design was applied where each participant was assigned to one of the proposal conditions (Welcome, ID&V or Transaction). The dependent variables explored were task performance (success/failure) and the navigational route (length and strategy employed) when selecting a service option at the main menu.

3. Results

The results analysis presented in this section was based on data entries from participants (*n*=114) who had managed to successfully complete all their four phone calls to the service[5].

3.7. Dialogue flow-chart and prompts for the overdraft request dialogue

This section describes the dialogue for the overdraft request. An overview flow-chart is presented in figure 3 and associated system prompts (for each error level of the main menu stages) are presented in figure 4.

Initial prompts (marked 'error=0' in figure 4) are played by the system which then waits for the caller to respond. Examples of caller responses are italicised in figure 3. Diamond shapes in figure 3 indicate routes through the dialogue where an erroneous input (silence or out-of-grammar) has been detected by the recogniser engine. The procedure for dealing with errors is to play a message relating to the input problem (*"I'm sorry, I didn't hear anything"* for silences; *"I'm sorry, I didn't understand that"* for out-of-grammar utterances), increment the error counter, and then to play the next error level menu prompt. Then, after three failed attempts at giving a valid input, the caller is offered to be transferred to a human operator.

[5] A further three participant data sets had been excluded due to call breakouts (which are transfers to a human agent caused by caller errors) occurring before the participant had a chance to attempt the overdraft task in the fourth call.

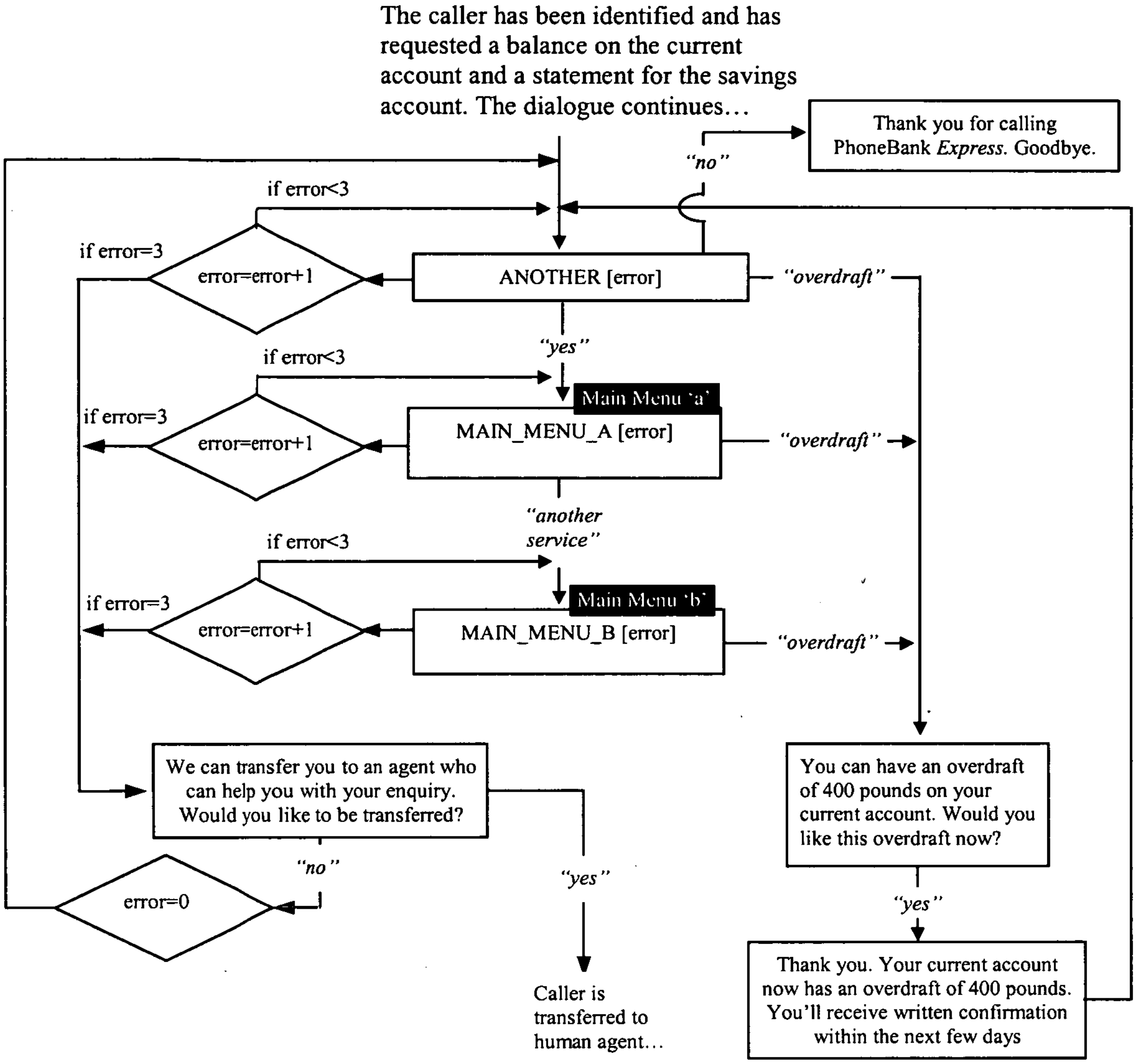

The caller has been identified and has requested a balance on the current account and a statement for the savings account. The dialogue continues…
Thank you for calling PhoneBank *Express*. Goodbye.
"no"
if error<3
if error=3
error=error+1
ANOTHER [error]
"overdraft"
"yes"
if error<3
Main Menu 'a'
if error=3
error=error+1
MAIN_MENU_A [error]
"overdraft"
"another service"
if error<3
Main Menu 'b'
if error=3
error=error+1
MAIN_MENU_B [error]
"overdraft"
We can transfer you to an agent who can help you with your enquiry. Would you like to be transferred?
"no"
error=0
"yes"
Caller is transferred to human agent…
You can have an overdraft of 400 pounds on your current account. Would you like this overdraft now?
"yes"
Thank you. Your current account now has an overdraft of 400 pounds. You'll receive written confirmation within the next few days

Figure 3. Flow-chart overview of the overdraft application dialogue. System messages appear in boxes and system flags (error level counter) appear in diamond shapes.

Prompt stage	Error Level	Prompt wording
MAIN_MENU_A	error=0	Please select balance, recent transactions or another service.
	error=1	Please say balance, recent transactions or another service.
	error=2	You can choose from balance, recent transactions, funds transfer, item search, order statement or change TIN. Please say the name of the service you would like or say help for further details.
MAIN_MENU_B	error=0	In addition you can select funds transfer, item search, order statement, or change TIN. Which service would you like?
	error=1	Please say funds transfer, item search, order statement or change TIN.
	error=2	You can choose from balance, recent transactions, funds transfer, item search, order statement or change TIN. Please say the name of the service you would like or say help for further details.
HELP (played if caller requests "help" as specified in MAIN MENU 'A' and 'B')	first time help requested	At this point you can get the balance on your account by saying balance, hear a list of the latest transactions by saying recent transactions, transfer money between your own accounts by saying funds transfer, search for a specific item on your account by saying item search, request an account statement through the post by saying order statement or change your secret telephone identification number by saying change TIN. Please select one of these options or say help for further details.
	second time help requested	If you would like to use your telephone keypad, for balance, press 1; for funds transfer, press 3; for order statement, press 4; for item search, press 5; for recent transactions, press 6; for change TIN, press 8. Which service would you like?
ANOTHER	error=0	Would you like another service?
	error=1	Would you like another service?
	error=2	You can either say yes or press 1, or say no or press 9. Would you like another service?

Figure 4. Prompts (with three error levels) for the main menu dialogue.

3.8. Task completion

Task completion rates were based on system log data and required that: (1) the current account balance had been played; (2) a statement had been ordered for the savings account; and (3) an overdraft had been requested. Overall, task completion rates for both the balance request and statement order were high (>90%, table 2). In the third phone call to the service, participants experienced a system-initiated overdraft proposal in one of three locations in the dialogue. In the fourth call to the service, participants were instructed (in addition to the balance and order statement task) to apply for an overdraft. Only 63% of participants successfully completed this request by saying "overdraft" at the main menu (table 2).

	Call 1	Call 2	Call 3	Call 4
Balance	112 (98.2%)	113 (99.1%)	113 (99.1%)	114 (100%)
Order Statement	103 (90.4%)	106 (93.0%)	109 (95.6%)	110 (96.5%)
Overdraft Request	-	-	-	72 (63.2%)

Table 2. Task completion success rates for each of the four phone calls to the automated service.

The remaining results analyses in this section focus on issues raised with regards to the overdraft task completion rates. Fisher's exact tests have been used for the statistical analyses throughout this section.

The first question is whether the location of the system-initiated overdraft proposal had an impact on participants' propensity to complete the overdraft task. Results showed that the overdraft task success rates for participants were: 71.8% (n=28) in the 'Welcome' proposal group, 56.8% (n=21) in the 'ID&V' proposal group, and 60.5% (n=23) in the 'Transaction' proposal group. However, this difference was not statistically significant (n=114, p=0.363).

Further analyses showed that male participants (figure 5) had a significantly higher overdraft task completion success rate overall than did female participants (n=114, p=0.006). 78% (n=39) of male participants managed to successfully select overdraft at the main menu; for female participants this figure was 51.5% (n=33). There were no significant differences for overdraft task completion rates based on age group.

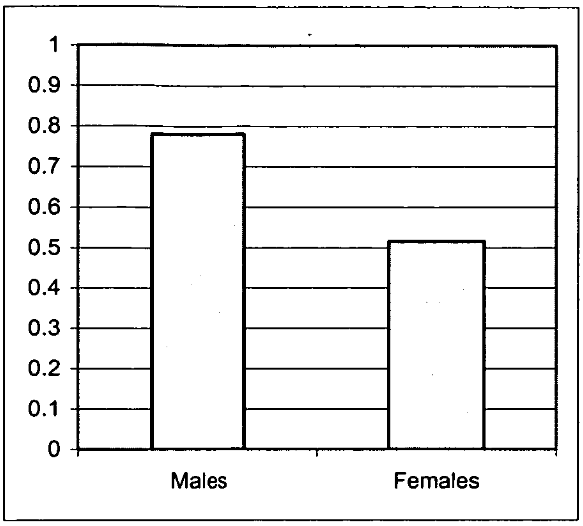


Figure 5. Overdraft completion ratio for males and females.

In each of the four phone calls to the service, participants tended to carry out the tasks in the order prescribed on the task sheet, i.e. balance first and then order statement. Once the participants had completed the balance request they were faced with the prompt "would you like another service?". At this point in the dialogue the order statement task could be carried out in three different ways. The participant could volunteer the "order statement" command and thus bypass the menu listings altogether. Alternatively, the participant could instead answer "yes" causing the service to loop back to the Main Menu 'a' prompt (figure 1): "please select balance, recent transactions or another service". Again, here participants could volunteer "order statement" or say "another service" in order to hear the Main Menu 'b' half of the menu listing. And finally, participants could then say "order statement" after waiting to hear the option being listed in Main Menu 'b'. Based on at which point in the dialogue "order statement" had been uttered, participants were grouped according to their propensity to 'volunteer' input information. Those who had said "order statement" before hearing it listed in the Main Menu 'b' in at least one of their four phone calls were labelled 'volunteered' (n=39); the remaining participants were labelled 'not volunteered' (n=75).

Figure 6 shows the overdraft task completion ratio, analysed in terms of participants' volunteering behaviour. The overdraft task success ratio for the 'volunteered' participants (n=33, 85%) was significantly higher than for the 'not volunteered' participants (n=39, 52%) who had said "order statement" only after hearing it listed in the Main Menu 'b' in all of their four calls (n=114, p=0.001)[6]. Only six participants (15.4%) in the volunteer statement group failed to complete the overdraft task (all of them having listened to both Main Menu 'a' and 'b').

[6] Similar statistical findings were obtained when using volunteering behaviour from the first two phone calls only (n=114, p=0.002). This indicates that the system-initiated proposal had little or no impact on participants' volunteering behaviour.

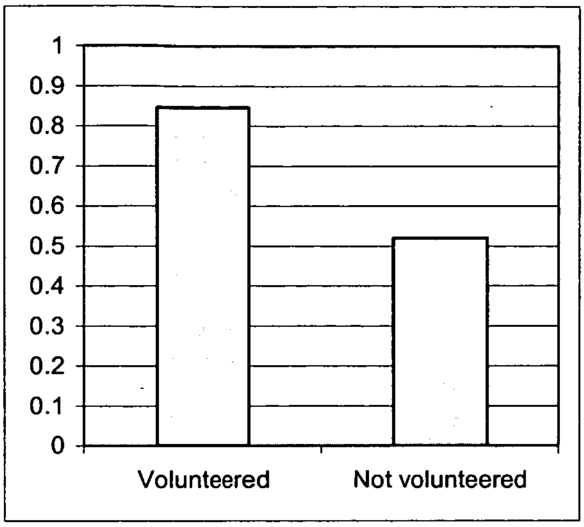


Figure 6. Overdraft completion ratio, based on whether participants had volunteered "order statement" at least once during their four phone calls.

The findings above suggest that volunteering behaviour and gender have a significant impact on participants ability to complete the overdraft task. The question therefore arises whether or not there are any significant variations in the propensity to volunteer input between males and females. Results showed that 42.0% (*n*=21) of males and 28.1% (*n*=18) of females volunteered statement in at least one of their phone calls, however, this difference was not strong enough to produce a statistically significant effect (*n*=114, *p*=0.164).

Figure 7 shows the overdraft completion ratio for males and females, split according to statement volunteering behaviour. Within the male participant group, volunteering behaviour did not have a significant effect on overdraft task completion (*n*=50, *p*=0.319); some 86% (*n*=18) of the male volunteered subset (*n*=21) completed the overdraft request compared with 72% (*n*=21) in the male not volunteered subset (*n*=29). In contrast, volunteering behaviour in the female participant group had a strongly significant impact on overdraft task completion (*n*=64, *p*=0.002); some 83% (*n*=15) of the female volunteered subset (*n*=18) completed the overdraft request compared with 39% (*n*=18) in the female not volunteered subset (*n*=46).

As described in the Participants section 2.2, about half of participants (*n*=50) stated that they had prior experience of using an automated telephone banking system for their personal banking. Of these, 58.0% (*n*=29) of participants with previous telephone banking exposure completed the overdraft, compared to 68.3% (*n*=41) of participants who had no previous exposure (non-significant, *n*=110, *p*=0.321). In terms of volunteering behaviour, 30.0% (*n*=15) of participants with previous exposure volunteered statement at least once during their four phone calls, compared with 36.7% (*n*=22) of those with no previous exposure.

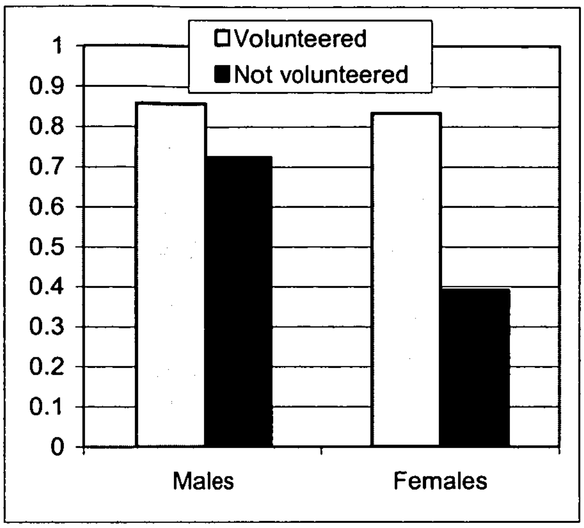


Figure 7. Overdraft task completion ratio for males and females participants, split according to whether or not they had volunteered "order statement" during at least one of their four phone calls.

3.9. De-briefing interview responses

A post-experiment interview was conducted in which the researcher asked those participants (*n*=42) who had not said "overdraft" during their fourth phone call the following question: *"Why did you not apply for the overdraft?"*. 40 participants responded to this question. One participant stated he had made a mistake and had answered "no" to the system prompt "would you like another service?". Out of the remaining 39 participants, the majority (79.5%) stated that they did not say "overdraft" as there was no such option in the main menu. Examples of actual participant responses from this group were: "I didn't get the option in the main menu" and "I expected the list of options to include an overdraft option".

The comments from the remaining eight participants (20.5%) reflected interpretation problems beyond the issue of the main menu listing. Four of these participants stated that they had not completed the overdraft application because they did not understand or remember how to: "Couldn't remember how to, was expecting to select it from the menu"; "Didn't understand how to apply for the overdraft. It was not obvious from the menu how to go about doing it"; "Wasn't anything on the menu regarding overdraft. Didn't know what the keyword was to get information on overdraft, so I stayed silent"; and "Couldn't remember how to. Don't think I heard overdraft message in the menu". Their comments indicate that they were aware that the option should be selectable at the menu but that they lacked the precise instructions for how to accomplish this. Another participant stated that: "I didn't get the overdraft message after the balance and couldn't remember what the third call message said about how to get an overdraft. Tried another service option but it didn't help". The remaining three participants expressed more general confusion: "It was confusing"; "I had forgotten what to say"; and "Didn't understand it. Got transferred to an operator".

3.10. Navigational strategy and route length

At each stage in the dialogue, system log files were used to record information about participants' interaction with the service. The log files contained details of the system prompt played, the error level, the recognition result returned by the system and whether the participant responded with voice or touch-tone button input. This data was then used to obtain information about each participant's navigational route through the system.

This section is concerned with participants' navigational route through the service up to the point of succeeding (or failing) to complete the overdraft request task. For this purpose, only the log data after the participant had completed the balance and statement order tasks in the fourth phone call will be considered (this is the part of the automated service which is outlined in the flow-chart in figure 3). Thus, the analyses described in this section start at the dialogue stage where the caller is faced with the system prompt "would you like another service?" and the task at hand at this point is to apply for an overdraft on the current account.

Participant actual responses for the overdraft request task
At this point in the dialogue, seven out of the 114 participants requested an overdraft: four of these participants said "overdraft", while three participants also specified the name of the account (e.g. "overdraft on my current account"). All seven participants had volunteered "statement" in at least one of their four phone calls to the service. One further participant made a mistake at this point in the dialogue and answered "no" which consequently ended the phone call; the remaining 106 participants responded "yes".

The "yes" response subsequently triggered the system prompt "Please select balance, recent transactions or another service" (Main Menu 'a' in figure 3). At this point in the dialogue, 26 participants said "overdraft": the majority of these (73%) had volunteered the "statement" keyword in at least one of their phone calls (two participants in the volunteer group had also stated the name of the account). The remaining 80 participants responded with "another service" which triggered Main Menu 'b': "In addition you can select funds transfer, item search, order statement or change TIN. Which service would you like?". Participant responses to this prompt are detailed in table 3 below.

	Volunteered statement at least once (n=13)		**Never volunteered statement (n=67)**	
Recognition result	**completed overdraft**	**failed overdraft**	**completed overdraft**	**failed overdraft**
"overdraft"	3	-	18	-
Other menu item	0	1	3	2
Invalid utterance	3	2	4	9
Silence	1	3	7	24
TOTAL	**7**	**6**	**32**	**35**

Table 3. Participant responses, after hearing all service options in the menu (Main Menu 'a' and'b').

In total, 21 participants requested "overdraft" at this point, while a further six participants requested other items from the main menu options (five said "item search" and one "balance"). A further 18 participants said something which was categorised as invalid utterances; of these 13 utterances were rejected by the system (85% of these utterances were of the category "none" or "none of these"), three utterances were misrecognised, and there were two cases of "another service". The remaining 35 participants remained silent at this point; silences constituted a significant part (46.3%) of the responses given by participants who had not volunteered "statement" in any of their phone calls to the service.

Navigational route length

This section details the total navigational route length for participants up to the point of either succeeding in completing the overdraft request (figure 8), or failing to do so (figure 9). The route length data complements the utterance analysis from the previous section by exploring the total number of turn-taking iterations involved in completing or failing/giving up on the overdraft task. For the purpose of this analysis, each system-prompt-user-response sequence was treated as one unit and received a score of '1' for navigational route length. The main reason for looking at the total route length (as opposed to actual path through the dialogue) is to reduce the number of route permutations for participants who persisted with looping back to the menu or trying alternative service options in order to find the overdraft. To recapitulate, only responses following the completion of the balance and statement order tasks will be included in the analyses, at the point where the caller is prompted with "would you like another service?" and the task at hand is to apply for an overdraft. Responses to this system prompt are represented by the route length of '1' in figure 8 and figure 9.

Figure 8 represents the route length by participants who completed the overdraft task, split according to whether or not they had volunteered statement in any of their calls. A route length of '1' means that participants responded "overdraft" to the system prompt "would you like another service?". A route length score of '2' means the participant had said "yes" to "would you like another service?", and then said "overdraft" at the Main Menu 'a' stage. Participants with a route length of '3' had said "another service" at Main Menu 'a' and then "overdraft" at main menu 'b'. Most of the participants in group '4' had, before saying "overdraft", a silence or reject and therefore had to go through an extra prompt-response sequence in the error recovery. The majority of participants with scores of '5' or over tried at least one other service before saying "overdraft". For comparison, navigational routes for participants who failed to complete the overdraft task are also included in figure 8. There was a tendency for participants in the volunteered statement group to request the overdraft before they had heard all the options in the main menu.

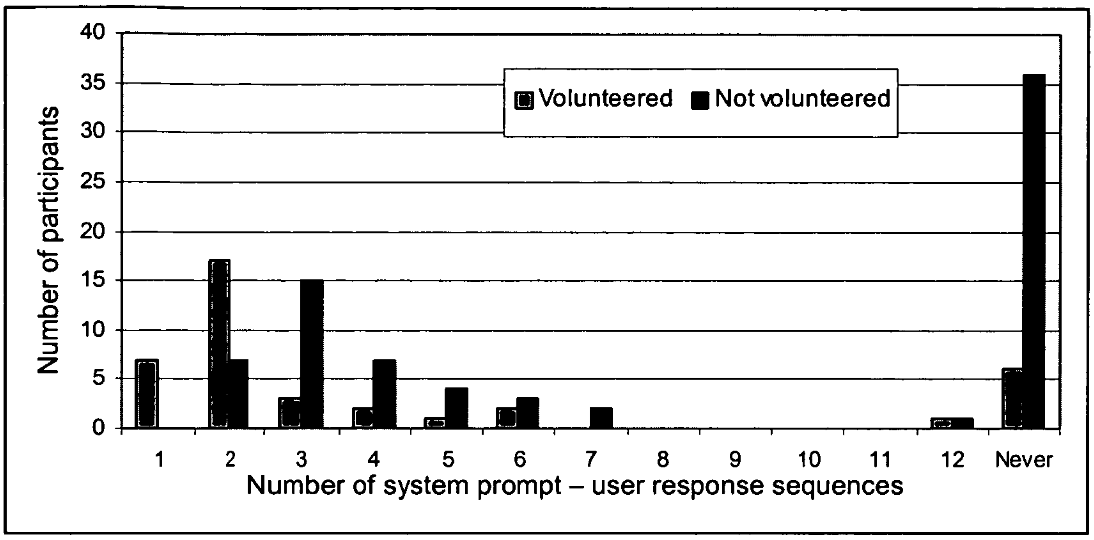


Figure 8. The length of the navigational route for participants who succeeded in completing the overdraft task, split according to their volunteering behaviour in the statement task. For comparison, participants who did not complete the overdraft task are included in the 'Never' category far right.

The same criterion for calculating the route length described in the previous section was used in figure 9, although instead of saying "overdraft" these participants had put the phone down or their call was transferred to a human agent. All but one participant (who had made the mistake of answering "no" in response to "would you like another service?") chose at least to hear both Main Menu 'a' and 'b' before their call ended. At the point of having listened to all the menu options, the majority of participants stayed silent or tried out the "item search" service option listed in Main Menu 'b'. In the real service, the item search option enables customer to search for a transaction on their account, either by giving the amount or cheque number. Participants, however, had not received priming about the functionality of any of the service options and so their assumption that 'item search' might help them with their task was a reasonable conclusion.

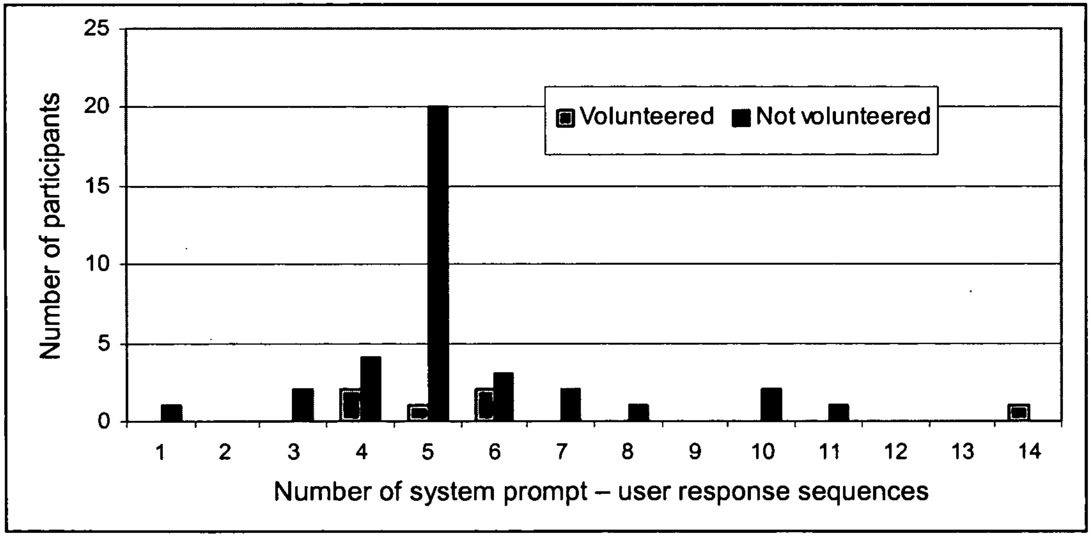


Figure 9. The length of the navigational route for participants who failed to complete the overdraft task, split according to their volunteering behaviour in the statement task. Their phone call either ended by hanging up, or by getting transferred to a human operator (call breakout).

4. Discussion

Today's speech-enabled automated telephone services typically rely on menu listings as a means of letting callers know about the range of options that they may choose from. An issue pertaining to the development and maintenance of such automated services is where to suitably introduce new or less frequently requested options, an area which has not yet been fully addressed in the current literature. The solution proposed in the current study involves the use of hidden options which are made accessible, but not explicitly listed, at the main menu; system-initiated proposals are then deployed in the dialogue to inform callers about the availability of the new service option. The purpose of this dialogue strategy is to try to encourage users to proactively request services, rather than to passively select options from the main menu listing.

The effectiveness and future implementation of such system-initiated proposals rely on the user's ability to successfully locate and select the hidden menu option. In order to assess this, an experiment was devised in which a new hidden 'overdraft' option was introduced into the dialogue of an already existing menu-driven automated telephone banking service in one of three locations: at the Welcome, ID&V and Transaction stage. The key finding from this experiment was that a significant proportion of the participant cohort (36.8%) failed to obtain an overdraft; in contrast, overall task completion rates for the menu-listed options balance and order statement were high (>90%). This leads to the conclusion that the system-initiated proposal --- in its present form --- is unsuitable as a method for introducing new, hidden, menu options into the dialogue of a mass-market automated telephone service. There were no significant differences in overdraft task completion with regards to the location of the proposal.

The main reason for participants failing to obtain the overdraft appears to emanate from their procedural and declarative knowledge of the automated service (i.e. how it works and how to operate it) --- commonly referred to as the user's mental model. A number of factors contribute to shaping this mental model: previous use of the service, experience of using other similar applications, information obtained from user guides, knowledge about how speech recognition technology works and so on. The user's model may not always be consistent with the system's conceptual model (Wærn 1993), an issue observable in the current experiment with the concept of hidden menus. The de-briefing interview revealed that the dominant interpretation of how to operate the automated service was to select an option from the main menu; and participants did not say "overdraft" because there was no such option in the menu to choose from. Furthermore, the navigational route length indicated that participants did not simply hang up once they had determined that the overdraft option was not included in the main menu: most participants looped through the main menu more than once before giving up. Essentially, these participants knew what they had to say, and where in the dialogue they should say it, but were prevented from doing so due to their failing to extrapolate beyond their ascribed strategy of selecting options from the menu. In fact, their conviction was so strong that they did not even consider attempting to just say "overdraft". These findings suggest that hidden menu options introduce dissonance in the user's mental model of the service.

This finding about the user's mental model of the service is interesting, but hardly surprising, considering that the concept of menu selection is enforced in most of today's automated telephone services. A closer examination of the navigational path length for participants (n=72) who had managed to complete the overdraft task further suggests that the concept of menu selection featured more prominently overall and that the hidden menu option therefore caused confusion: 23 of these participants (31.9%) had at least one extra turn-taking iteration (silence or chose an alternative option) after having listened to all options of the two main menu halves. What, then, enabled some participants to succeed with the overdraft request in the dialogue where others failed? The experimental results suggest that there were two major factors at play: volunteering behaviour and sex differences.

The first of these two --- volunteering behaviour --- derived from participants' propensity to say "order statement" during the first half of the main menu (before they had heard the option listed). The experimental results showed that participants who volunteered the statement keyword also were more successful at completing the overdraft task, which in turn suggests that they were less likely to abide by the main menu listings. The reason behind this kind of volunteering behaviour is not clear, but the results indicate that their approach may have been 'accidental' rather than strategic in that they may not have been fully aware that they did not always wait for the option to be listed in the main menu before selecting it; this is supported by the finding that six participants who volunteered "statement" in at least one of their calls did not volunteer the "overdraft" keyword. One possible explanation behind this behaviour is that these participants' responses were triggered by the task instructions presented on the priming sheet, rather than by the menu options in the system prompts. Based on the experimental findings, it is not possible to determine whether or not the volunteering behaviour would feature in real-world use of the automated service.

The second contributing factor to overdraft success/failure in the experiment concerns user gender differences: the fact that, unexpectedly, significantly more male than female participants managed to complete the overdraft task. The impact of gender difference on overdraft task success was particularly noticeable when taking into account statement volunteering behaviour: the link between volunteering behaviour and overdraft task completion was stronger in the female participant group then in the male. In other words, there seemed to be a trait prominent more so in the male participant group which facilitated the completion of the overdraft request. These findings are interesting and suggest that male and female users may have different abilities or traits, or that they may adopt different problem solving strategies when operating automated telephone services. Psychometric analyses were outside the scope of the current research and therefore the underlying factors responsible for this behaviour could not be identified.

4.11. Limitation and future study

The findings from the current research have laid bare some interesting facts in terms of the usability of menu-driven automated telephone services, but have also generated further questions regarding issues surrounding individual users' abilities to operate such applications. The purpose of this section is to highlight these issues and to guide a research direction for further study.

The first issue to be considered is, in light of the low rate of successful overdraft requests, the use of hidden menu options as a means for introducing new options into the service. A perhaps obvious solution to the user's problem (male or female) of obtaining an overdraft would be to simply add an overdraft option to the main menu listing. Using this approach, the purpose of the system-initiated proposal would then be reduced to notifying callers that a new service option has been added, or to advertise particular features pertaining to the new option that may be of interest to the caller. This approach was employed in a follow-up experiment (unpublished results) which employed a similar setup, and resulted in all participants successfully obtaining an overdraft. However, adding service options to the main menu in this way is not an ideal solution as it will render listings longer and more cluttered. An alternative method would be to keep the overdraft option hidden, and instead revisit the contents and wording of the system-initiated proposal. A suitable approach would be to use the proposal to resolve the primary cause for participants' failing to obtain an overdraft, i.e. redress their erroneous assumption that only service options listed in the main menu can be selected. In this case, the proposal would be used to 'educate' users that --- although the overdraft option is not explicitly listed in the main menu --- it can still be accessed by saying the "overdraft" keyword. Alternatively, in the dialogue explored in the current research, the Main Menu 'b' listing could be reworded to instruct users about extra options; for example, a more open-ended "or just tell me which service you are interested in" may encourage callers to volunteer input rather than select it from the menu of 'core' options. There is opportunity for further research in this area, as each such new proposal design needs to be assessed in terms of its impact on service usability, user attitudes and task completion.

Secondly, the current experiment results identified that the caller's mental model of the service (i.e. strictly selecting options from a menu) prevented them from completing the overdraft request. The underlying factors behind such mental models were, however, not fully addressed in the experiment. It is suggested that future studies into hidden menu options should also probe the mental models amongst participants who *succeeded* in obtaining the hidden option to complement the data from participants who failed to do so. How did these participants reason? Were participants, who volunteered input rather than selecting it from the menu, aware that they did not always wait to hear the option first? What exactly prompted them to request the overdraft? Furthermore, the demographic data captured in the current experiment was limited; more detailed information about the type of telephone applications participants experienced in real-life and the frequency of use would need to be obtained in order to provide a full account of the potential impact that habituation effects may have on users' mental models. Such data could then be used to explore if there is a link between habituation effects and user strategies for operating the automated services (e.g. the tendency to volunteer input or select options from menus).

Thirdly, and perhaps most prominently, are questions raised by the gender differences revealed in the current experiment. In order to understand the full significance of the impact of gender differences on the usability of automated services, and in order to establish whether or not there is a link between cognitive ability and task completion, further research is necessary. There are standardised psychometric tools available that can be used to measure differences in cognitive abilities (e.g. spatial ability, verbal ability, problem solving skills) and which have been applied in previous research into automated telephone services. For example, Foster et al. 1998 noted that the level of user's spatial ability affected how participants rated the mode of data entry (touch-tone buttons, isolated and connected word recognition for speech input). Goldstein et al. 1999 found that task completion times varied according to spatial ability and prompt strategy: a guided (menu) prompt strategy seemed to better suit participants with low spatial ability and the open ("what do you want to do now") prompt strategy was more suitable for participants with high spatial ability. Research into differences in cognitive ability between men and women has met with some controversy (raising issues whether the differences are due to biological or social factors) and the significance of such findings has often been disputed. However, it is generally accepted that men, on average, are better at a range of spatial skills than are women; whereas women are better at some tasks requiring memory for the location of objects (Kimura 1999). Evaluation of participants' cognitive skills was outside the scope of the current experiment. It is suggested that future experiments, which aim to explore participants' ability to locate hidden menu options, also include an element of psychometric evaluation.

The findings from this research (mental models, volunteering behaviour, menu navigation, sex differences) extend beyond the domain of banking dialogues and are relevant to the design of a range of menu-driven automated telephone services. In speech-enabled applications, menus may facilitate the interaction for novice or infrequent users by promoting a step-by-step interaction, but can also render the interaction in speech-driven applications unnecessarily stilted and long. The challenge to designers of such applications is to strike a balance between restricting the user inputs and at the same time conveying to the user a conceptual model which allows them to fully exploit the strength and flexibility of the speech recognition technology. With the automated telephone services becoming ever more ubiquitous in society, and with the increased application of speech recognition in such services, this is a research domain well worth exploring.

Acknowledgements

This work was made possible by the generous support of Lloyds TSB Bank PLC and Nuance Communications Inc.

5. Appendix

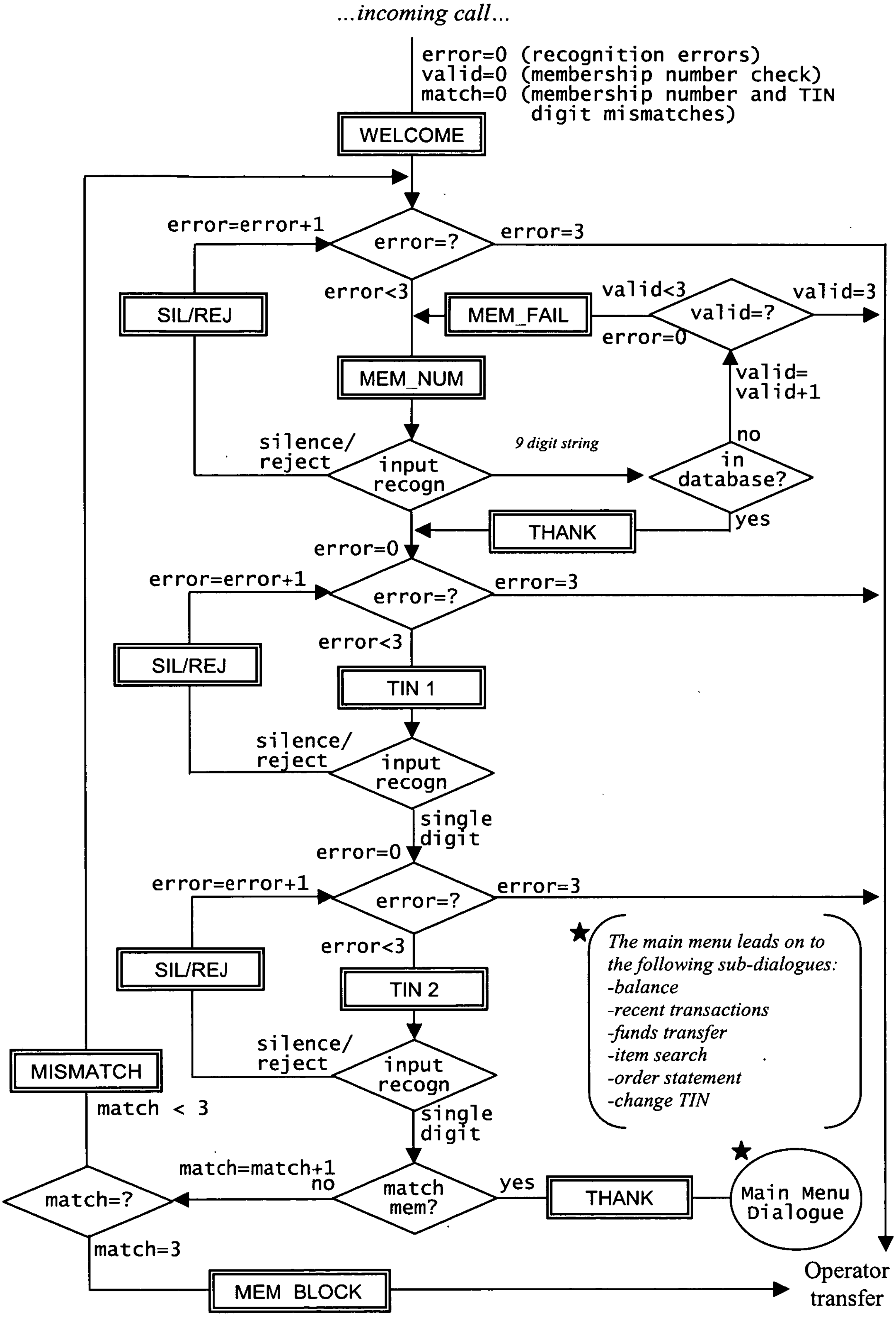


Figure 10. Flow-chart for the identification process in the banking dialogue. Diamond shapes show system checks and recognition stages. System prompts are enclosed in rectangles.

Prompt stage	Error Level	Prompt wording
WELCOME	n/a	Welcome to PhoneBank Express.
MEM_NUM	error=0	Please give your membership number now.
	error=1	Please give your nine digit membership number now.
	error=2	Your membership number has nine digits and is printed on your membership card. You can either say the nine digits or enter them on your telephone keypad. Please give your membership number now.
MEM_FAIL	valid=1, 2	I'm sorry, that membership number doesn't match our records.
TIN 1	error=0	Please give the [X] digit of your secret TIN now.
	error=1	Please just give the [X] digit of your secret TIN now.
	error=2	You can either say the digit or enter it on your telephone keypad. Please just give the [X] digit of your secret TIN now.
TIN 2	error=0	...and the [Y] digit.
	error=1	Please just give the [Y] digit of your secret TIN now.
	error=2	You can either say the digit or enter it on your telephone keypad. Please just give the [Y] digit of your secret TIN now.
MISMATCH	match=1	I'm sorry, there seems to be a problem so I'll need to ask you for your membership number again. You can either say the digits or enter them using your telephone keypad.
	match=2	I'm sorry, this is the second time I've been unable to match your responses against our records. Please call back after checking your details.
MEM_BLOCK	match=3	I'm sorry, as this is the third time I've been unable to match your responses against our records, we're suspending your use of the service for your own security. We'll send you a new personal identification number shortly so you can call the registration line and re-register to use the service again.
SIL/REJ	silence	I'm sorry, I didn't hear anything.
	reject	I'm sorry, I didn't understand that.
THANK	n/a	Thank you.

Figure 11. Prompt recordings used in the identification and verification process.

6. References

BERNSEN, N.O., DYBKJÆR, H., DYBKJÆR, L., 1996, Principles for the design of cooperative spoken human-machine dialogue. Proceedings of the International Conference on Spoken Language Processing (ICSLP) '96, Philadelphia, October 1996, pp. 729-732.

COHEN, M.H., GIANGOLA, J.P., BALOGH, J., 2004, *Voice user interface design* (Addison-Wesley), ISBN 0-321-18576-5.

Dialogues 2000 Report, 1997, Navigation in structured and unstructured menus. Experiment Series Report No. 6, Centre for Communication Interface Research, University of Edinburgh, UK.

FOSTER, J.C., MCINNES. F.R., JACK, M.A., LOVE, S., DUTTON, R.T., NAIRN, I.A., WHITE, L.S., 1998, An experimental evaluation of preferences for data entry method in automated telephone services. *In Behaviour & Information Technology*, 1998, Vol. 17, No. 2, pp. 82-92.

GOLDSTEIN, M., BRETAN, I., SALLNÄS, E.-L., BJÖRK, H., 1999, Navigational abilities in audial voice-controlled dialogue structures. *Behaviour & Information Technology*, 1999, Vol. 18, No. 2, pp. 83-95.

HORNSTEIN, T., 1994, Telephone voice interfaces on the cheap. Proceedings of the UBILAB '94 Conference, Zurich, pp. 134-146, Universitätsverlag Konstanz, Konstanz, September 1994, pp. 134-147.

KARIS, D., DOBROTH, K.M., 1991, Automating services with speech recognition over the public switched telephone network: human factors considerations. *IEEE Journal on Selected Areas in Communication*, Vol. 9, No. 4, May 1991, pp. 574-585.

KIMURA, D., 1999, *Sex and Cognition*, MIT Press, 1999, ISBN 0-262-11236-1.

KOTELLY, B., 2003, *The art and business of speech recognition: creating the noble voice* (Addison-Wesley), ISBN 0-321-15492-4.

MCINNES, F.R. , NAIRN, I.A., ATTWATER, D.J., JACK, M.A., 1999, Effects of prompt style on user responses to an automated banking service using word-spotting. *BT Technology Journal*, vol.17, no.1, January 1999, pp.160-171.

PREECE, J., ROGERS, Y., SHARP, H., 2002, *Interaction design: beyond human-computer interaction* (John Wiley & Sons Inc.), ISBN 0-471-49278-7.

SHEEDER, T., BALOGH, J., 2003, Say it like you mean it: priming for structure in caller responses to a spoken dialogue system. *International Journal of Speech Technology*, 6, pp. 103-111.

TOMKO, S., ROSENFELD, R., 2004, Shaping spoken input in user-initiative systems. Proceedings of the International Conference on Spoken Language Processing (ICSLP-Interspeech) '04, pp. 2825-2828.

VANHOUCKE, V., NEELEY, W. L., MORTATI, M., SLOAN, M. J., NASS, C., 2001, Effects of prompt style when navigating through structured data. Proceedings of INTERACT 2001, Eighth IFIP TC.13 Conference on Human Computer Interaction, IOS Press, Tokyo, Japan, 2001, pp. 530-536.

WALKER, M.A., FROMER, J., DI FABBRIZIO, G., MESTEL, C., HINDLE, D., 1998, What can I say?: evaluating a spoken language interface to email. Proceedings of ACM CHI 98, Conference on Human Factors in Computing Systems, pp. 582-589.

WÆRN, Y., 1993, Varieties of learning to use computer tools. *Computers in Human Behavior*, Vol.9, pp. 323-339, 1993.

WEINSCHENK, S, BARKER, D.T., 2000, *Designing effective speech interfaces* (John Wiley & Sons Inc.), ISBN 0-471-37545-4.

WILKIE, J., JACK, M. A., LITTLEWOOD, P., 2002, Design of system-initiated digressive proposals for automated banking dialogues. Proceedings of the International Conference on Spoken Language Processing (ICSLP) '02", Denver, Colorado, pp. 1493-1496.

WILLIAMS, J.D., SHAW, A., PIANO, L., ABT, M., 2003a, Evaluating real callers' reactions to Open and Directed strategy prompts. Applied Voice Input/Output Society Speech Developers Conference/SpeechTEK Spring EXPO, March/April 2003, San Jose, California.

WILLIAMS, J.D., SHAW, A., PIANO, L., ABT, M., 2003b, Preference, perception, and task completion of Open, Menu-based, and Directed prompts for call routing: a case study. Eurospeech, September 2003, Geneva, Switzerland.

YANKELOVICH, N., 1996, How do users know what to say? *ACM Interactions*, Vol. 3, No. 6, November/December 1996, pp. 32-43.

Available online at www.sciencedirect.com
SCIENCE DIRECT•

Int. J. Human-Computer Studies 62 (2005) 41–71

International Journal of
Human-Computer
Studies

www.elsevier.com/locate/ijhcs

System-initiated digressive proposals in automated human–computer telephone dialogues: the use of contrasting politeness strategies

J. Wilkie[a,*], M.A. Jack[a], P.J. Littlewood[b]

[a]Centre for Communication Interface Research, School of Engineering and Electronics, The University of Edinburgh, Edinburgh EH9 3JL, UK
[b]Lloyds TSB Bank Plc, Canons House, Canons Way, Bristol BS99 7LB, UK

Received 19 September 2003; received in revised form 2 July 2004; accepted 6 August 2004
Communicated by S. Brave

Abstract

System-initiated digressive proposals may be used to introduce new and unexpected information into automated telephone services. These digressions may be viewed as particularly pronounced forms of unsolicited interruptions as they contain information not directly related to the caller's intended activity. In human–human conversation, interruptions are considered to be speech acts which intrinsically threaten both the positive and negative face wants of the addressee and conversants adopt specific verbal strategies to mitigate the negative impact of their interruptions. A question therefore arises whether the introduction of face-redressive expressions, based on human–human conversational strategies, into the design of system-initiated proposals in automated services can mitigate the negative impact of the interruptions. A usability experiment was conducted to examine the effectiveness of three contrasting politeness strategies for system-initiated digressions in a mass-market telephone banking dialogue using speech recognition technology. Participants ($N = 111$) experienced these proposals while using the automated service to perform banking tasks. Results indicated

*Corresponding author. Tel.: +44-131-6502801; fax: +44-131-6502784.
E-mail address: jenny.wilkie@ccir.ed.ac.uk (J. Wilkie).

doi:10.1016/j.ijhcs.2004.08.001

that all these system-initiated digressions—irrespective of politeness strategy employed—had a negative impact on the user attitudes towards the service. This paper reports these results and explores participants' perceptions of the politeness styles and registers employed in the system-initiated proposals.

Keywords: Politeness theory; Natural language interface; Dialogue design; Dialogue evaluation; Usability; Automated telephone banking; System-initiated digression; Interruption

1. Introduction

Speech recognition technology is increasingly used in the mass-market domain of self-service telephone applications. Compared to their push-button counterparts, applications which use spoken language input offer users a more natural and flexible way of interacting with a computer-based system. However, the system messages and the turn-taking in these speech operated applications often still resemble those found in push-button operated services in that they follow a rigid prompt-response sequence where the input options are presented to users in the form of vocal menus and explicit instructions about what to say. The dialogue between the human user and the automated service in such applications typically follows a pre-defined script involving a fixed turn-taking structure (the computer prompts then the user responds) and valid user responses are restricted by the capabilities of the speech recognition grammar. Users of these mass-market applications can expect a controlled and predictable interaction with the computer in a dialogue that does not change between phone calls.

Mass-market automated services are primarily designed to handle task-driven conversations within a narrow topic domain, such as flight information, banking account transactions or cinema bookings. The user of such services typically expects the interaction to be restricted to the chosen topic and task at hand and that the computer will cooperate fully to complete the goal of the call. Fixed turn-taking, goal-driven, prompt-response interaction has become the conventional way of designing automated self-service telephone applications. It is not common practice for an automated service to initiate an interruption or launch into new topics, a fact which may explain why such dialogue behaviour remains largely unexplored in the current literature for human–computer spoken interaction. The possibility of deploying system-initiated digressions in human–computer conversation raises new and interesting dialogue engineering issues regarding the design, usability and acceptability of such applications.

The research described in this paper explores how system-initiated digressive proposals may be used to disseminate unsolicited financial information in a speech-driven automated telephone banking service. These proposals work by interrupting the user and suspending the regular dialogue turn-taking for the duration of the informational message. The key issue examined in this research is how politeness strategies (considered an important factor in the choice of vocabulary in human–human dialogue interruptions) may be employed to influence the impact of such system-initiated digressive proposals on user attitudes.

2. Motivation

Whilst the capability of speech recognition technology is continuously improving, the need still exists for the system messages (or prompts) to be designed to control for the range of user inputs that can be accepted by the computer (Bernsen et al., 1997). Much of the dialogue design for mass-market automated telephone services is centered around making the interaction fit the technology at hand, relying on explicit instructions and error recovery strategies in order to guide users as to what to say, how to say it and when to speak. In mass-market applications, menus in the form of explicit list selections are usually employed as a method for informing users (especially novice users) about the range of services available to them. Once such automated services have been designed, implemented and launched, changes to the dialogue turn-taking, menus or prompts can be costly, can result in customer objections and are rarely made. As a result, less frequently demanded information and transactional services are usually excluded from the dialogue and there is no straight-forward way of adding or deleting service options from menus once the application is up and running.

There are a number of reasons for looking beyond conventional menu-based dialogue design to explore alternative and more flexible means of offering users access to services through an automated application. For example, an enterprise may want to introduce new informational or transactional services that may not be considered in the initial application design under normal circumstances, such as access to services which are infrequently requested, short-term offers or product promotions, but at the same time avoid adding these as options to menus which may become unnecessarily long and complex. Furthermore, successful take-up of an automated service may result in the enterprise losing opportunities to advise customers about relevant products or services. This may involve the use of a "logical link", such as when a specific transaction on a customer's account is used to trigger advice on a relevant service or product.

One solution for adding new options in the service is to introduce a short system-initiated informational message within the dialogue structure with the intention of disseminating new information relevant to a particular customer at a specific point in time during their use of the automated service. This system-initiated message could simply consist of a brief prompt which may or may not be followed by a short dialogue (e.g. requesting a yes/no response) enabling the user to pursue or decline the offer immediately. The system-initiated message interrupts the regular turn-taking of the dialogue and, in doing so, impedes the human user from continuing with the flow of the call as anticipated. These messages may therefore be viewed as a particularly pronounced form of system-initiated digression since they are in effect unsolicited, unexpected and not directly related to the current topic or the prime goal of the call.

Successful strategies in this area could have important positive commercial implications.

3. Background

Dialogue engineering for speech-driven mass-market applications is mainly concerned with development issues relating to the technology at hand, such as whether to use voice recordings or text-to-speech for system prompts; whether to allow callers to barge-in during system prompts; whether to use open or closed prompt styles (Hone and Baber, 1999); whether to use isolated word recognition or allow for more fluent speech; and whether to allow universal commands such as "cancel" or "exit". The design principles for automated dialogues and research into voice interactive services described in the recent literature (Balentine and Morgan, 2001; Gardner-Bonneu, 2001) offer broad coverage of design aspects. However, they offer little in terms of guidelines on how to implement system-initiated interruptions and digressions in speech-driven applications: the area addressed in this paper. Studies of digression (often referred to as "out-of-turn interaction" or "unsolicited reporting" (Allen et al., 1999)) that may be found in human–computer spoken interaction research have focused mainly on providing models for handling *user*-initiated digressions (Haller, 1994; Narayanan et al., 2000; Ramakrishnan et al., 2002) which occur when the user supplies extra or out-of-turn information in response to system prompts. This new or extra information supplied by the user is however normally related to the overall goal of their participation in the conversation.

Previous research by the authors (Wilkie et al., 2002) identified two key dialogue engineering issues for the design of system-initiated digressions in human–computer dialogues: the *location* of the proposal in the application call-flow; and the dialogue *turn-taking strategy* employed for delivering the proposal information. In order to address these design issues, two experiments were devised in which system-initiated proposals were introduced in the call-flow of a speech-driven automated telephone banking service (Wilkie et al., 2002). These proposals informed users that they could apply for an overdraft facility on their account through the automated service by saying "overdraft" at the menu of services (the main menu). The wording employed in the proposal was low-key, short and terse in order to be consistent with the register employed in the rest of the automated service. In order to soften the impact of the interruption, phrases such as "*You might like to know that...*" were used in the opening statement of the proposal.

In order to assess the impact of the digressions in that research, participants' attitudes towards the telephone banking service were measured before and after they were subjected to this additional overdraft message. Participants were not forewarned about the pending interruption, nor did they receive any experiment priming to create a potential interest in the overdraft product. Results from that research revealed that participants' attitudes towards the usability of the service remained unaffected by the delivery of the overdraft proposal during their phone call. Additionally, measurements of participants' attitudes towards the digressive dialogue itself suggested that there were no overall strong indications that one particular location or turn-taking strategy was more favoured than the other. These findings suggest that automated telephone banking dialogues can be successfully augmented using system-initiated digressive proposals. The results from that research serve as a point of departure for this current investigation into other aspects of system-initiated digressions in automated dialogues, where the role of politeness strategies and the effect of the stylistic manner employed when interrupting are investigated.

4. Politeness in human–human interaction

Politeness in human communication has received much attention in the field of pragmatics and sociolinguistics over the past two decades and has mainly been focused on how communicative strategies are employed in order to promote and maintain social harmony[1] in human–human interaction. One of the most influential theories of politeness is that developed by Brown and Levinson (1987). Their politeness theory is based on the notion that each individual has positive and negative "face wants" and that these are ascribed by all (rational) interactants to themselves and to one another in any social interactive situation. Brown and Levinson define the two face wants as (1987, p. 61):

> Negative Face: the desire to be un-impeded in one's actions, the basic claim to territories, personal preserves, rights to non-distraction—i.e. freedom of action and freedom from imposition.
> Positive Face: the desire (in some respects) to be approved of, the positive consistent self-image or "personality" claimed by interactants (crucially including the desire that this self-image be appreciated and approved of).

Any utterance or action in a communicative situation can be seen as potentially threatening to the positive or negative face of either of the interactants and, consequently, expressions of politeness are normally used as mitigations aimed at redressing this threat. Although Brown and Levinson give examples of how the speaker's *own* positive and negative face wants may be at risk,[2] much of their politeness theory is primarily focused on the explicit strategies used by a speaker to avoid damaging the addressee's face wants (Chen, 2001). The speaker uses politeness expressions to indicate that no face threat is intended or desired and to convey that the addressee's face wants are recognized and approved of by the speaker.

Brown and Levinson calculate the relative "seriousness" of a face-threatening act based on three "social dimensions" (1987, p. 74). These are: the relative power of the addressee over the speaker; the social distance between the speaker and the

[1] For an account of impoliteness strategies see Culpeper (1996).

[2] For example, expressing thanks or making an excuse are damaging to the speaker's negative face. Admissions of guilt or non-control of emotions (laughter or tears) are examples of damage to a speaker's positive face.

addressee; and the ranking of the imposition involved in doing the face-threatening act. Brown and Levinson point out that each of these dimensions is context-sensitive, meaning that the relationship between two individuals (such as the relative power of a manager over an employee) may be inverted under certain circumstances. Depending on the seriousness and social setting for the face-threatening act, a number of options are presented to the speaker on how to redress a potential face threat. First of all, the speaker has the option of not performing the act at all and could therefore theoretically avoid damaging the face of the addressee altogether. However, if the speaker decides to go ahead with the face-threatening act, Brown and Levinson identify a taxonomy of politeness which includes four principal categories of expression strategies: (1) doing the act without redressive action (baldly), (2) using positive face-redress, (3) using negative face-redress, or (4) doing the act off-record. The "off-record" strategy attempts to minimize the face threat by creating uncertainty as to the existence of the face-threatening act itself, e.g. by using ambiguous or vague expressions, or by using hints such as "it's cold in here" (implying "shut the window").

To carry out an act "baldly", without redress, involves doing it in the most direct, clear, unambiguous and concise way possible (Brown and Levinson, 1987, p. 69). The speaker may use the bald strategy when there is no fear of retribution by the addressee (e.g. in the interest of urgency or efficiency, e.g. "watch out!"); where the danger to the addressee's face is very small (such as in proposals and requests); or where the speaker is considerably superior in power to the addressee.

Face-redressive politeness strategies are used when there is a perceived potential threat in an utterance to either the positive or negative (or both) face wants of the addressee. Utterances that are considered threatening to the *negative* face wants of the addressee will include: ordering the addressee to do something, making an offer which may incur debt for the addressee and expressions of strong emotions towards the addressee. Negative face-redressive strategies are characterized by formality and distancing. It is such forms of "negative politeness" that are conventionally associated with politeness in everyday language, such as "excuse me" and "thank you", as these relate to the imposition itself. *Positive* face-redress, on the other hand, widens the sphere of politeness to include the appreciation of the addressee's wants in general or to the expression of similarity between speaker's and addressee's wants. Threats to the addressee's positive face wants are caused by, e.g. bringing bad news about the addressee, expressing disapproval or raising emotionally divisive topics. The positive face-redress strategy is characterized by "intimate" language behaviour and makes reference to a close interdependent social relationship between the interactants. For example, the speaker might use in-group identity markers (hey buddy) or show intensified interest in the addressee's wants (your hair looks *great*).

Some face-threatening acts, such as interruptions,[3] are considered to be intrinsically threatening to *both* the negative and positive face wants of the addressee

[3] Other face-threatening acts considered to intrinsically threaten both the negative and positive face wants of the addressee are complaints, threats, strong expressions of emotions and requests for personal information.

(1987, p. 67). By Brown and Levinson's definition, an interruption constitutes a threat to the negative face wants of the addressee because it infringes to some degree on the addressee's right to non-distraction and desire to be un-impeded in their actions. Interruptions also pose a threat to the addressee's positive face wants by implying that the person who interrupts ignores or does not care about the addressee's feelings and wants.

5. Relevance to human–computer interaction

Politeness is undoubtedly an important aspect of *human–human* conversation, but little prior work has been undertaken to investigate how relevant it is to *human–computer* dialogues. What are the conversational rules or social dimensions that govern the use of politeness registers in dialogues where one of the interactants is a computer? Can existing politeness theories be expanded to encompass human–computer interaction? If so, what politeness strategies should the computer (in the capacity of the speaker) be endowed with and how are the resulting politeness expressions received by the human user?

People's interactions with computers (and other media) are fundamentally social (Nass et al., 1994; Reeves and Nass, 1996; Nass and Moon, 2000). This view is founded on the notion that the human brain has evolved to respond and relate socially to human-like entities in our surroundings and that this innate reaction is almost impossible to overcome—even in situations where humans interact with a supposedly non-social entity such as the computer. This propensity for humans to relate socially to media has been explored in a series of controlled experiments (Reeves and Nass, 1996; Nass and Moon, 2000). The results showed that users applied gender stereotypes to computers; they identified with computer agents sharing their ethnicity; and they were more attracted to agent characteristics (submissive/dominant) that were similar to their own personality. The authors also concluded that users apply "over-learned" social rules to computers, such as politeness: experimental results showed that participants gave a significantly more positive evaluation of a computer's performance when questioned directly by the computer itself compared to when questioned by a different computer or through pen and paper questionnaires. This would indicate that politeness is an important factor in human–computer interaction. However, the work on politeness in human–computer interaction carried out by Nass and colleagues has been centered around how humans behave politely towards computers, rather than investigating how humans respond to a computer that tries to portray polite behaviour.

The experimental results obtained by Nass and colleagues strongly suggest that human users have a subconscious tendency to apply deeply rooted social rules to interactions with computers in the same way as they do when interacting with other fellow humans. These social rules seem to relate to our innate disposition and cultural upbringing. But how do users react to a computer that blatantly attempts to exploit these social rules? Fogg and Nass (1997) explored the effects of employing computer-initiated flattery when giving feedback to users in a text-based guessing game application. Experimental results showed that flattering feedback (compared to the generic feedback condition) had a positive effect on a number of aspects of the interaction. For example, the flattery increased participants' feelings of power; made them more positive towards their own and the computer's performance; and made them enjoy the interaction more.

Colón et al. (2001) studied the use of politeness in interruptions in a graphical library search engine interface. These interruptions involved on-screen error text messages (resulting from either system errors or user errors) that were presented with or without politeness (courtesy). The messages were deployed in the library application and evaluated in a controlled experiment. The two main findings from the experiment were: firstly, the interruption performed by the computer interface had a detrimental effect upon the user perception of the interaction with the computer (the participants judged the interaction as being less friendly, less motivating and less beneficial). Secondly, it was found that politeness strategies had no effect on minimizing participants' negative reaction towards the interruption.

The idea of treating the computer as a social entity and endowing it with emotive qualities such as politeness may be considered to be controversial given the fact that the computer does not have any real understanding about the effect its behaviour may have on its dialogue partners. Some user interface designers are opposed to the idea of anthropomorphizing computers and stress that users should be discouraged from thinking that computers may have human-like abilities (Shneiderman, 1988, 1993, 1998). This position derives from the point of view that human relationships are rarely a good model for designing effective human–computer user interfaces and that the primary goal for interface design should be predictable and controllable interaction (Shneiderman, 2000). McFarlane (1998), in his work on interruptions of the visual display in human–computer interaction, concludes simply that politeness is an irrelevant topic for the design of user interfaces as computers do not have "face" and people do not have face-wants relative to their computers. MacFarlane therefore suggests that the "bald" strategy is adequate for these purposes and should be employed.

Much of the research effort into the social aspects of human–computer interaction has been focused on the visual screen interface, which is operated by keyboard and mouse. The human–computer interaction that takes place through speech over a unimodal telephone channel is different from the visual interface and possibly even more sensitive to linguistic and social effects. The use of language in a user interface (and the use of speech in particular) is considered one of the most likely characteristics of technology that prompt a social response (Nass in Anderson, 2000, p. 95). Automated telephone services rely on speech output and the characteristics of the voice (such as the pitch, register and tone) carry sensitive information about personality and identity of the speaker. For example, Boyce (2000) compared a number of contrasting voice personalities which ranged from "from butler to hip youth" in a voicemail system and found that users reacted differently to these extremes. Some participants "loved" the butler personality whereas others found "him" annoying; the voice personalities that exhibited least extreme speaker characteristics caused fewer negative reactions from users (but also resulted in fewer really strong positive reactions). Furthermore, the social interaction appears to be enforced further by the use of speech recognition technology in that it is not uncommon for users of speech-driven telephone applications to answer politely "yes please" or "no thank you" in response to system prompts.

6. The politeness experiment

6.1. Introduction

To explore issues in politeness with automated telephone dialogues, a controlled usability experiment was conducted in which participants ($N=111$) experienced system-initiated digressions while they performed banking tasks using an automated telephone banking service. The system-initiated digressive proposals explored in this research explicitly stated in the opening phrase that the proposal constituted an interruption. This forthright method is likely to be perceived by users as more intrusive compared to the more low-key "*you might like to know*" opening phrases used in previous research (Wilkie et al., 2002, summarized in Section 3), but may however better serve to alert users to the ensuing information by capturing their attention. The purpose of making deliberate digressive interruptions in the current experiment was to explore if politeness strategies for human–human interaction (as defined by Brown and Levinson, 1987) could be employed to mitigate the adverse effects of these dialogue intrusions.

The experiment had four conditions based on the prompt register applied in the proposal: (1) Positive face-redress, (2) Negative face-redress (3) Bald (no face-redress) and (4) A no-proposal control condition. Participant attitudes towards the proposals were assessed, both in terms of the relative politeness of the proposal strategy in the context of the automated banking service (main experiment) and, secondly, the absolute politeness (Leech, 1983) associated with the face-threatening act, independent of dialogue context. The absolute politeness was established by allowing participants to listen to each individual proposal over computer speakers after they had completed the main experiment.

It is anticipated that user attitudes to system-initiated digressions will vary according to the relevance of the information to the user's specific situation. Determining what is, or is not, relevant to an individual caller is a complex matter involving modelling of the caller's intentions, wants, needs and goals: most of which are in the mind of the user and not accessible to the computer system. The research reported here does not address issues relevant to defining the business criteria or user models for deciding whether or not to make a proposal to a particular caller on a particular occasion; rather, it assumes that the decision to deploy the digression has already been taken. This approach is comparable with real life situations in which a (human) call centre agent reviews a customer's accounts and decides to approach the customer with a product offer, which may or may not be related to the original purpose of the phone call. The agent perceives a potential need for the product but has little insight into the customer's general financial situation or needs, external to the details at hand.

Results from previous experiments (Wilkie et al., 2002) have already asserted that system-initiated digressions can successfully be deployed in the automated service without relying on complex user models. In the current experiment, the information contained in the system-initiated digression was chosen on the grounds of being financially beneficial and applicable to the majority of callers: i.e. the new "On-line Saver" account offers a higher interest rate than the accounts that the customer currently holds.

6.2. Participants

Participants were recruited from the general public in Edinburgh. Although some participants had used an automated telephone service for their personal banking, no previous experience of automated telephone banking was required in order to take part in the experiment. In total, 111 complete participant data sets were attained and used in the statistical analysis (Table 1).

6.3. Experiment procedure

Participants were told that they would use an automated telephone banking service to perform some banking transactions. For ethical and data protection reasons, no personal data were used at any point in the experiment. Participants were presented with a sheet of paper containing their fictitious persona to be used throughout the experiment: a membership number, a 6-digit personal telephone identification number (TIN) and details of "their" two (fictitious) accounts (a current account and a savings account). Prior to the first call to the automated service, participants were given a task sheet instructing them to find out and make a written note of the balance of "their" current account. Between phone calls participants were asked to imagine that "a few days had gone by" and that they were then to call the service to check their balance again. In total, participants made five phone calls to the automated service (the No-proposal control group made only three phone calls). The experiment proceeded in a number of clearly defined stages which are detailed below.

6.3.1. Two phone calls (No-proposal)

Each participant was asked to make two phone calls to the automated service (without any savings proposals being made at this stage). This procedure allowed all

Table 1
Participant demographics

	Age 18–35	Age 36–49	Age 50+	Total
Female	28 (25.3%)	18 (16.2%)	17 (15.3%)	63 (56.8%)
Male	23 (20.7%)	15 (13.5%)	10 (9.0%)	48 (43.2%)
Total	51 (46.0%)	33 (29.7%)	27 (24.3%)	111 (100%)

of the participants to become familiar with the service functionality and their persona details. Following these two phone calls, the participants completed a questionnaire (referred to here as USAB1) to establish their attitude towards the usability of the service for later comparison after having experienced the proposal in their third phone call. The "USAB" questionnaire contents are detailed in Section 6.6.2.

6.3.2. *Third phone call (with system-initiated proposal)*

In the third phone call to the service, three-quarters of the participants experienced one of three randomly selected contrasting system-initiated digressions (Positive, Negative or Bald) while carrying out their banking enquiry (the No-proposal control group simply used the same banking service they had experienced in the previous two calls). By design, in order to avoid pre-empting participant reactions to the digression, no mention of savings proposals had been made up to this point in the experiment. Following this third phone call, all participants completed a second service usability questionnaire (USAB2). An additional questionnaire (P.ROP1) was also administered to participants who had experienced a proposal delivery during their phone call. The "PROP" questionnaire (detailed in Section 6.6.3) was targeted at user attitudes towards the interrupting digression itself (as moderated by the politeness strategy).

6.3.3. *Two phone calls (additional proposals)*

Participants (excluding those in the No-proposal group) were asked to make two additional phone calls to the service. These phone calls allowed participants to experience the remaining two face-redressive strategies in a controlled randomized order. Participant attitudes were assessed following each of these phone calls (PROP2 and PROP3). For these final two phone calls, only the questionnaire concerning attitudes towards the proposal itself (PROP) was used.

6.3.4. *Listening session*

After all of their phone calls to the service had been completed, participants[4] in the No-proposal control group listened to each of the three savings proposals over computer speakers. This was done in order to obtain a measure (manipulation check) of the absolute politeness of the proposals when abstracted from the context of the telephone banking dialogue. The results of the listening session data analysis revealed that the contrasting registers used in the proposals carried significantly discernable information regarding politeness attributes and that these findings were in-line with Brown and Levinson's theory. In summary: (1) the Negative face-redressive strategy was perceived to be most polite, apologetic and respectful and the

[4] All 111 participants took part in this session, however, only the data from the No-proposal control group ($N = 25$) were used in the analysis so as to avoid participants' responses being influenced by their experience of the proposal delivery in the context of the automated service.

speaker using the strategy was perceived to be most tactful, professional and caring; (2) the Positive face-redressive strategy was perceived to be least polite, formal, to the point and respectful and the speaker using the strategy was perceived to be least tactful and professional; and (3) the Bald strategy was perceived to be the most unapologetic, formal and to the point. For further details about these findings, see Appendix A.

6.3.5. Exit interview

A structured interview was then conducted containing questions relating to participants' perceptions and preferences of the politeness strategies used. Finally, participant details such as age and familiarity with automated telephone banking services were recorded. Participants then received an honorarium payment of £20.

6.4. The automated banking dialogue

The automated telephone banking service used in this research was modelled on an existing real-world application which provides customers with access to personal account information (e.g. balance information or recent transactions) and enables them to perform a number of banking transactions such as funds transfers. The service enables users to employ spoken natural language input by allowing for some extraneous speech (*Can I have…, please*) and the possibility of giving multiple pieces of information at the main menu (e.g. "*I'd like the balance of my current account, please*").

The application dialogue is outlined in the simplified flow-chart in Fig. 1, showing the system prompts and user responses for the identification and verification stage, followed by a balance enquiry.

The prompt style in the dialogue is terse, with politeness expressions limited to "please", "thank you" and "I'm sorry". Speech input is promoted throughout the service dialogue; each dialogue stage features a three error-level recovery where push-button options are mentioned in the third level (error recovery) prompt. The banking dialogue was implemented using commercially available speech recognition software. Prompts were recorded using a recording artist (female) who has a Southern British English accent.

6.5. Design of the digressive dialogue

The resulting three proposal variants have the following basic design criteria in common: they start out with an explicit interruption (mitigated by contrasting politeness strategies); they point out the financial benefits to the customer; they give details of restrictions that apply (that transfers to and from On-line Saver accounts can only be done via telephone or Internet banking); and, finally, they allow interested customers to pursue the offer immediately by engaging them in a "yes/no" (follow-on) dialogue. If the customer answers "no" at this point the service continues

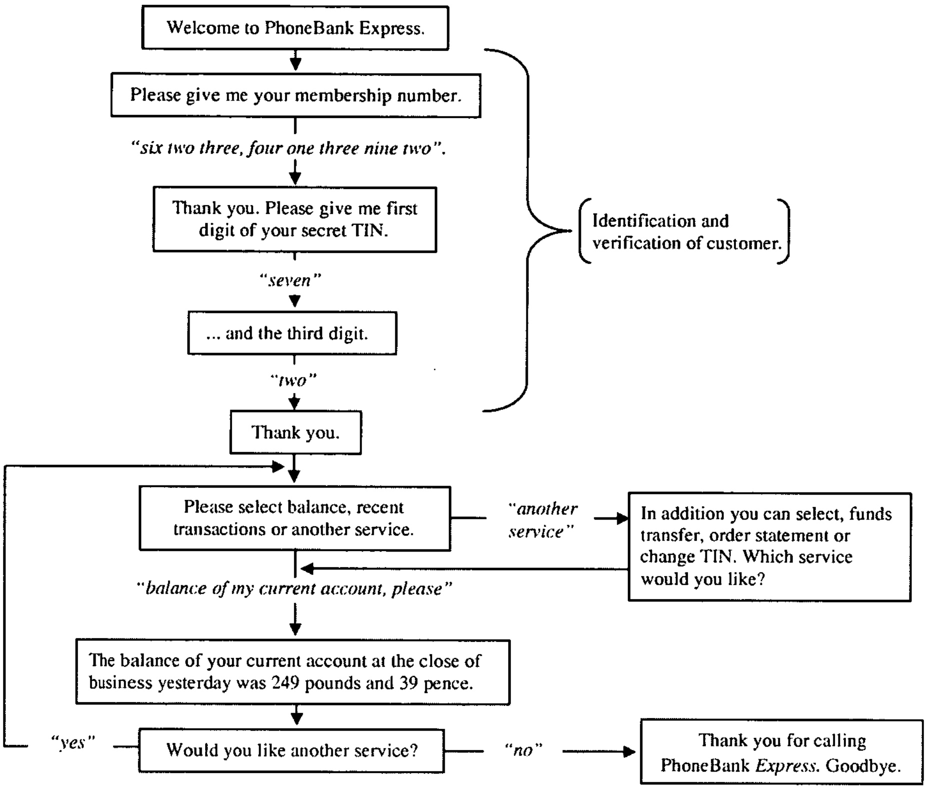

Welcome to PhoneBank Express.
Please give me your membership number.
"six two three, four one three nine two".
Thank you. Please give me first digit of your secret TIN.
"seven"
... and the third digit.
"two"
Thank you.
Identification and verification of customer.
Please select balance, recent transactions or another service.
"another service"
In addition you can select, funds transfer, order statement or change TIN. Which service would you like?
"balance of my current account, please"
The balance of your current account at the close of business yesterday was 249 pounds and 39 pence.
"yes"
Would you like another service?
"no"
Thank you for calling PhoneBank Express. Goodbye.

Fig. 1. Overview of the banking dialogue. System messages appear in boxes, and user responses are italicised.

the dialogue with "Would you like another service?". Participants who answer "yes" to the proposal hear the following message (note that the application procedure was simplified for experimental purposes):

> "Thank you, your new On-line Saver account will be available from tomorrow."

For the purpose of the experiment, the system-initiated proposals were deployed immediately after a caller had been uniquely identified (after obtaining the two secret TIN digits, just before the prompt with the menu options: "*Please select balance*..." in Fig. 1). In real-world use, this location would ensure that only eligible customers were offered the savings proposal and that the proposal could be monitored such that it would be offered only once to each customer. The wordings for each of the three contrasting styles of proposals are detailed in the following sections.

6.5.1. *Positive face-redress register*

Brown and Levinson's theory states that threats to the addressee's positive face (through an interruption) are mitigated by using expressions of solidarity, informality and familiarity. Examples of positive face-redress are, exaggerating the interest in the addressee; sympathizing with the addressee; and avoiding disagreement. In the current experiment, the Positive face-redress was realized by the following linguistic devices (Brown and Levinson, 1987, pp. 101–129):

Being optimistic: "*I know you won't mind...*"
Informality: "*...cutting in...*"
Intensifying interest with the addressee: "*...special information for you...*"
Exaggerating approval with addressee: "*...make your growing savings grow even more*".
Presupposing common ground: "*we all want the best return possible...*"
Showing concern for the addressee's wants: "*with your interests in mind, I suggest...*"
Offering and promising: "*...an On-line Saver account that will give you better interest...*"
Giving or asking for reasons: "*why not set one up today!?*"

Proposal with Positive face-redress—(prompt recording 30 s long)

"I know you won't mind me cutting in with some special information for you about how to make your growing savings grow even more. We all want the best return possible from our savings. With your interests in mind, I suggest you open an "On-line Saver account" that will give you better interest than the accounts you've got just now. You can transfer money to and from an On-line Saver account through telephone or Internet banking. Why not set one up today! Do you want me to do that for you now?"

6.5.2. *Negative face-redress register*

Negative face-redress involves expressions of restraint, formality and distancing, such as being conventionally indirect, giving deference and apologizing. In the current experiment the negative face-redress was realized by the following linguistic contents (Brown and Levinson, 1987, pp. 129–211):

Apologizing: "*I'm very sorry to interrupt...*"
Stating the face-threatening act as a general rule: "*it is the bank's policy to notify...*"
Impersonalizing speaker and addressee: "*...notify customers how to...*"
Being indirect: "*we wish to inform you...*"
Giving deference: "*...as a valued customer...*"

Being pessimistic: "*you may therefore want to consider...*"
Going on record as not indebting addressee: "*we would be happy to...*"

Proposal with Negative face-redress—(prompt recording 31 s long)

"I'm very sorry to interrupt, but it is the bank's policy to notify customers about how to improve their savings returns. We wish to inform you, as a valued customer, that an "On-line Saver account" offers better interest than the accounts you hold at present. You may therefore want to consider opening an account of this type. Transfers to and from On-line Saver accounts are made through telephone or Internet banking. We would be happy to set up an On-line Saver account for you today. Would you like us to do that now?"

6.5.3. *Bald register (No face-redress)*

Undertaking a speech act without positive or negative face-redress is described by Brown and Levinson (1987) as performing the act "baldly". In contrast to the registers used to mitigate positive and negative face threats, the primary concern in the *Bald register* is to be direct and concise. The Bald register is applied under circumstances where the face threat can be ignored, in the interest of urgency and efficiency. The speaker might, e.g. feel that the information is so important or interesting to the addressee such that there is no need for a more convoluted expression. Alternatively, the speaker might be unconcerned about any imposition on behalf of the addressee. The Bald proposal in the experiment was stripped of any kinds of face-redress and started with: "*I'm interrupting to inform you about...*".

Bald proposal (No face-redress)—(prompt recording 18 s long)

"I'm interrupting to inform you about how to improve your savings returns. The "On-line Saver account" offers better interest than the accounts you have at present. You can transfer money to and from an On-line Saver account through telephone or Internet banking. Do you want to set up an On-line saver account now?"

6.6. *Usability evaluation*

6.6.1. *Aim*

The experimental research had two principal aims: (1) to establish whether the presence of a digressive interruption (as moderated by politeness strategy) influenced participant attitudes towards the usability of the automated service; and (2) to evaluate the effects of contrasting politeness strategies on participant attitudes towards the interruption itself (in the context of the automated service dialogue).

6.6.2. *Measurement of overall service usability*

The design of the usability questionnaire (referred to in this paper as USAB) followed standard practice (Likert, 1932) by using an equal number of negative and positive statements presented to the respondent in a randomized order. In this way the danger that the overall usability score could reflect the respondent's tendency to agree rather than disagree with the questionnaire statements (an effect known as "*response acquiescence set*") is removed. Respondents mark their opinion for each statement by ticking the appropriate box along 7-point Likert scales that range from "Strongly Agree" (1) to "Strongly Disagree" (7). Following reversal of the polarity of positive questionnaire statements, in this paper a score of 7 consistently indicates a strong positive attitude and 1 a strong negative attitude.

Previous research has identified key attributes required for evaluating the usability of automated telephone interfaces and for assessing the contributions to usability made by each of the attributes (Love et al., 1992) by means of written questionnaires. The usability questionnaire used in this research consisted of 20 statements that address a range of issues pertaining to human–computer telephone interaction: *cognitive issues* (level of concentration and degree of confusion), *the fluency and transparency of the service* (knowledge about what is expected, ease of use, degree of complication), *system performance* (reliability of service, efficiency of service, amount of improvement service is felt to require) and *system voice* (clarity of the voice, politeness of the service, friendliness of the service).

All participants ($N = 111$) completed the service usability questionnaire following two "practice" phone calls (USAB1) and then again after their first exposure to the dialogue which included the system-initiated proposal (USAB2). Comparisons of the mean scores from these two questionnaires were used to establish the impact of the system-initiated proposal on participant attitudes towards the usability of the service.

6.6.3. *Measurement of attitudes towards digressive proposals in the dialogue*

Participants' reactions towards the system-initiated digressions were evaluated using two different approaches. Firstly, the manipulation check (detailed in Appendix A) allowed control group participants to experience each proposal in isolation over computer speakers, in essence removing the proposals from the context of the dialogue and focusing the analysis on the qualitative aspects inherent in the contrasting politeness strategies employed.

The second assessment approach involved capturing participant reactions towards the proposal interruption itself (as moderated by politeness strategy) in the context of the automated service dialogue. For this purpose, a supplementary set of questionnaire statements (referred to as PROP here) was added to the USAB usability questionnaire. The proposal attributes included in the PROP questionnaire were: *relative disruptiveness* (whether the proposal was annoying, intrusive, distracting and interrupted the call too much), *face-redressive characteristics* (polite, friendly, formal, apologetic, patronizing, manipulative, caring for individual needs), *durational aspects* (length and long-windedness), *information quality* (helpfulness, efficiency, relevance of contents, appropriateness to context), *cognition* (ease of understanding) and *trust* (confidence in service, willingness to pursue the offer through the service). Participants ($N = 86$, excluding the No-proposal control group who did not experience a proposal in the context of the automated service) responded to these questionnaire statements (PROP1-3) following each exposure to a system-initiated proposal (which occurred in phone calls three, four and five during the experiment). The mean scores for these questionnaire items enabled direct comparisons of participants' attitudes towards the proposals based on the three contrasting politeness strategies employed.

7. Results

The results analysis was performed in two separate stages. Firstly, an assessment was made of the impact of the system-initiated proposal on participants' attitudes towards the automated service dialogue (USAB). This was achieved by comparing how participants ($N = 111$) rated the overall service usability before (USAB1) and after (USAB2) experiencing the first proposal delivery, which occurred in the third call to the service. The No-proposal (control) group used exactly the same service on all their three phone calls and did not experience a proposal in any of these calls.

The second analysis stage investigated participants' ($N = 86$) perceptions towards the system-initiated proposals specifically (PROP) and explored how attitudes towards the interruptions were affected by employing contrasting politeness strategies. The analysis compared participants' attitudes towards the contrasting proposal strategies based on their response data (PROP1) following the very first exposure to a proposal (in call three) and then, in a separate analysis, by pooling the mean scores (PROP1-3) following exposure to all three proposal variants (calls three, four and five).

In the analysis of the questionnaires, scales with participant responses were adjusted for polarity to ensure that all mean scores below 4 indicate a negative response to the statement, whereas values above 4 indicate a positive response.

7.1. *Usability attitudes towards the automated service dialogue*

The dependent measures used in these analyses were the mean responses to the questionnaire statements on service usability completed after the two familiarization calls (USAB1) and after the first exposure to the product proposal in the third call (USAB2).

The mean usability scores prior to and following the introduction of the unsolicited proposal for each individual experimental condition are shown in Table 2. Within each proposal group, repeated measures ANOVAs were carried out with age (3 levels) and gender as between-subject variables. Results showed that

Table 2
Mean usability scores based on proposal condition

Proposal condition	*N*	Mean usability before proposal	Mean usability after proposal	Statistical results
No proposal[a]	25	5.71	5.72	*df=1, F=.14, p=.710*
Positive face redress	29	5.56	4.94	*df=1, F=10.27, p=.004*
Negative face redress	28	5.79	5.10	*df=1, F=24.25, p=.000*
Bald (no face redress)	29	5.72	4.83	*df=1, F=28.06, p=.000*

[a]Control group.

there was no significant change in attitude for the No-proposal group before and after call three, but for each of the three proposal conditions there is a noticeable drop in the attitude towards the service following the introduction of a proposal delivery. The change in attitude for each of the three proposal groups in Table 2 was highly significant ($p<.01$).

A univariate ANOVA was carried out based on the mean score differences between the two questionnaires (USAB2-USAB1). The between-subject variables used in the first analysis were age (3 levels), gender and presence/absence of proposal (2 levels). Results showed that, when compared to the No-proposal control group, the overall drop in attitude for participants who experienced the presence of a proposal was significant [$df = 1$, $F = 15.05$, $p = .000$]. Analysis of mean score differences for individual questionnaire items revealed that proposal group participants found the service significantly ($p<.01$) more *frustrating* and *less enjoyable to use*, making it *less efficient* and *more in need of improvement.* Participants felt *more under stress*, *less in control* when using the service and they were *less happy about using the service again.* At a lower level of significance ($p<.05$), participants found the service with the proposal *more confusing*, *more complicated* and *less easy to use*. They felt *more flustered* when using the service, they had to *concentrate harder* and *knew less what to do*.

A further univariate ANOVA analysis was based solely on the mean score differences between the three proposal groups, with age (three levels), gender and proposal strategy (three levels) as the between-subject variables. The results showed that there was no significant difference overall between the three proposals [$df = 2$, $F = 1.85$, $p = .165$]. Further analyses revealed that only two questionnaire items were statistically significant ($p<.05$): *the service was easy to use* and *the service was reliable.*[5]

[5] Where the results of statistical tests such as *t*-tests or ANOVAs show only one or two significant differences in a set of 20 questionnaire items, these should not be relied on since it is a statistical fact that when a number of such tests are carried out there is a high probability that at least one at the 95% level will be a false positive.

Table 3
PROP questionnaire mean scores, based on proposal condition and call number

Proposal condition	PROP1 mean first proposal	PROP2 mean second proposal	PROP3 mean third proposal	Overall mean
Positive	4.02	3.90	3.57	3.83
Negative	3.81	4.20	4.14	4.06
Bald	3.72	4.50	4.45	4.22
Total mean	3.85	4.20	4.05	

In summary, the presence of a proposal in the dialogue had a significantly negative impact on service usability overall, while there were no significant differences in mean score differences between the three proposal groups. These results suggest that it was the presence—rather than the politeness strategy—of the proposal that had the major impact on attitudes toward the service usability in this experiment.

7.2. *Attitudes towards the digressive proposals*

Two analyses were carried out on the PROP questionnaire items, which were specifically aimed at capturing user attitudes towards the proposal interruption and the politeness strategy employed. Firstly, a univariate ANOVA was carried out on the responses after the participant's first exposure to the proposal dialogue (PROP1). Secondly, repeated measures ANOVA was carried out on the data after exposure to all three contrasting proposals (PROP1-3), where responses had been pooled according to the style of the proposal (Positive, Negative and Bald). The No-proposal control group did not experience any proposals and are therefore not included in these analyses. Mean scores for the PROP questionnaires are presented in Table 3.

7.2.1. *Analysis based on first exposure to a proposal*

The between-subject analysis was performed on participant responses (PROP1) to their first exposure to the proposal dialogue, which occurred during their third phone call to the service. This simulates the reactions from customers who encounter the (unsolicited) proposal for the first time during automated telephone banking. The results from the analysis show that there were no significant differences between proposal strategy groups overall. There were some differences in attitude for individual items in the questionnaire however, based on the sample size used in this experiment, these were not strong enough to produce statistically significant result.[6] Thus, it can be concluded that there were no differences in the way participants responded to the contrasting face-redressive strategies employed in the proposals based on first proposal exposure.

[6] In fact, the questionnaire item regarding *"the proposal interrupted the call too much"* showed a weakly significant difference. This result, on a single item in a 24-item questionnaire, could easily be due to chance.

The mean scores show that the general attitude towards any of the three proposals was negative; more than half of the questionnaire item scores fell below (or nearing) the neutral point 4 on the 7-point scale for all proposal strategies. Items that were aimed at eliciting the face-redressive characteristics of the proposal register received generally positive responses. In particular, the items relating to the *politeness* and *friendliness* achieved mean scores on, or above, 5 on the scale. In addition, with scores above neutral, participants did not seem to think that either proposal was *too apologetic*, *too formal* or *patronizing*. Questionnaire items that resulted in markedly negative responses (scores below 3) related to the disruptiveness of the proposal in the call: the proposal was perceived to be *distracting, intrusive, too long, annoying* and believed to *interrupt the call*.

In summary, results based on a participant's first exposure to the proposal indicate that there were no significant differences between politeness strategies employed, with regards to the sample size used in this analysis.

7.2.2. *Analysis of pooled response data*

Participant responses for all three questionnaires (PROP1-3) were pooled according to politeness strategy. The pooled-data approach has two main advantages: it increases sample size and enables the use of within-subject comparisons (which in turn reduces the unsystematic variability in the design and provides greater power to detect effects). The main disadvantage with the pooled-data approach is that it includes data from the second and third proposal calls where the proposal content no longer is new or unexpected (creating a learning effect). As a consequence, if participant responses to the first exposure are significantly different compared with subsequent exposures, then two different conditions—"proposal novice" and "proposal-aware" participant groups—are mixed in the results.

A repeated-measures ANOVA was carried out on the pooled data with one within-subjects variable (proposal strategy) and three between-subject variables: age (3 levels), gender and proposal order (6 levels, based on all possible controlled permutations of exposure). The analysis of the overall difference between participant mean scores (column labelled "Overall Mean" in Table 3) revealed an overall statistically significant difference for proposal strategy [$df = 2, F = 4.629, p = .012$]. Within-subject contrasts showed that this difference lay between the Positive face-redress proposal and the Bald strategy [$df = 2$, $F = 11.432$, $p = .001$].

The analysis also showed a significant interaction between proposal strategy and order group [$df = 10$, $F = 2.528$, $p = .009$], indicating that participant attitudes towards the proposals were confounded with one of the following: (a) order effect due to call number; (b) exposure to preceding proposals, or; (c) an interaction between call number order effect and the current proposal wording. Following this finding, the data were adjusted to compensate for the effect due to call number by subtracting the overall questionnaire mean for each call number (e.g. 3.85 for PROP1, Table 3) from individual participant mean scores within that proposal call. The new mean scores were then used in a re-run of the repeated-measures ANOVA. Results showed that that the significant effect of proposal strategy remained unchanged, but that the interaction between proposal strategy and order of exposure

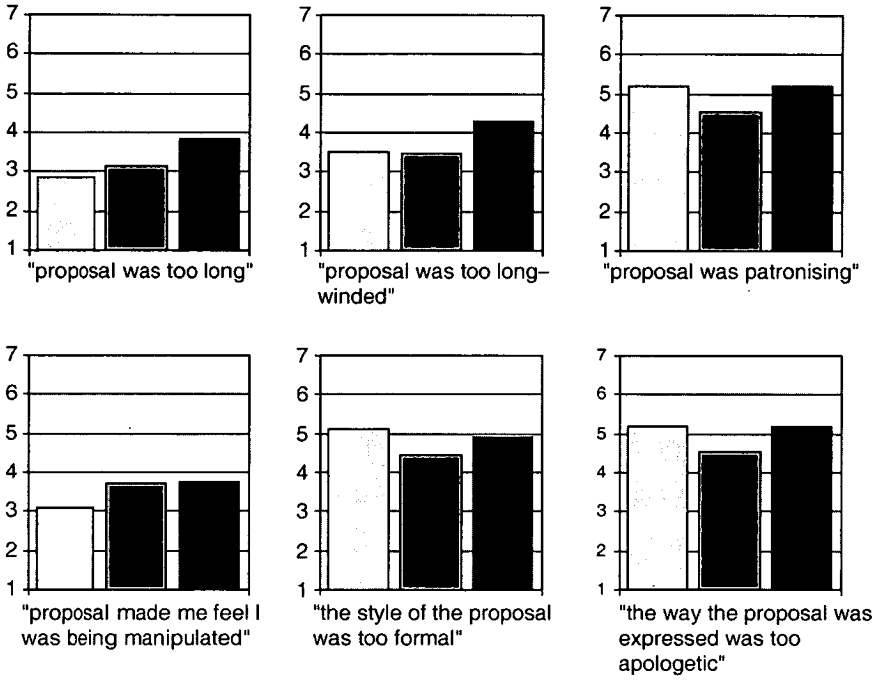


Fig. 2. Mean responses by condition. □ Positive; ▣ Negative; ■ Bald.

became much weaker and was no longer significant [$df = 8$, $F = 1.059$, $p = .240$].[7] These results support the theory that there are two simple effects present: a simple effect of proposal strategy regardless of previous exposure and a simple effect of call number regardless of strategy involved. To conclude, this suggests that there is a genuine effect of proposal strategy, applying in both the "novice" and the "proposal-aware" conditions, but it is not conclusive since these results were not reflected in the analysis of the data from the first proposal call.

Further ANOVAs, performed on individual statements in the unadjusted pooled data scores, revealed a number of attributes with highly significant differences between the three proposal styles, as illustrated graphically in Figs. 2 and 3 (mean scores and results from the statistical analysis are shown in Table 4). Note that higher mean scores indicate a more positive and supportive attitude towards the concepts conveyed by the Likert statements in the questionnaire. For example, the first of the charts in Fig. 2 reveals that the Bald proposal generated a more positive response regarding the proposal length compared to the Positive and Negative proposal strategies.

In addition to being favoured in terms of its shorter *length*, the Bald proposal was also perceived by participants to be significantly *less long-winded* compared to the

[7] The *F*-value and degrees of freedom here have been adjusted for the fact that the means used for compensation were estimated from the data.

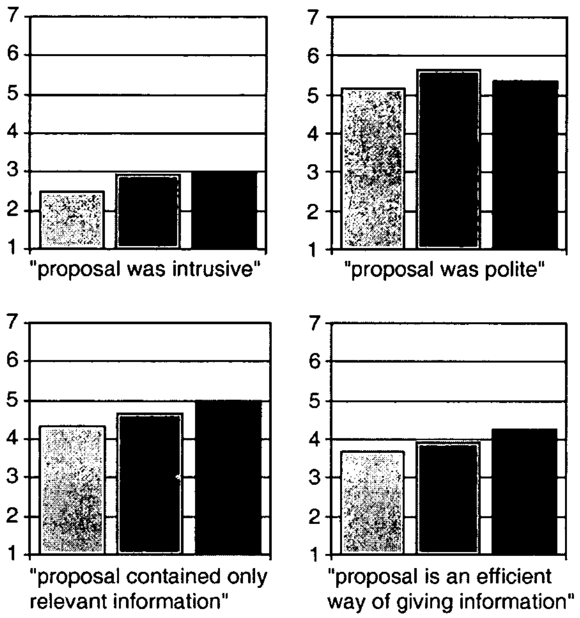


Fig. 3. Mean responses by condition. ▨ Positive; ▣ Negative; ■ Bald.

Positive and Negative face-redressive proposal strategies. This suggests that attributes such as the length and wordiness of a proposal have a strong impact on user attitudes towards system-initiated digressions.

Fig. 2 also highlights participants' reactions towards the face-redressive characteristics employed in the contrasting politeness strategies. The Positive face-redress proposal which relied on an informal and intimate register was found by participants to be *more manipulative* than the Negative face-redress and the Bald strategy. The Positive face-redress proposal was also perceived to be significantly more *patronizing* than the Bald strategy. The Negative face-redress was found to be significantly more *formal* and *too apologetic* when compared to the Bald strategy and the Positive face-redress.

Participant responses highlighted in Fig. 3 give some further indications to participants' perceived differences of the proposal strategies. In terms of the relative *intrusiveness* there was a significant difference in attitude between the Positive face-redressive proposal and the Bald strategy, the Positive face-redress being perceived to be *more intrusive*. The Negative face-redressive proposal was rated most *polite* of the three, with the difference between the Positive and Negative proposal strategies approaching highly significant ($p = .012$). Noticeably, all three proposal strategies received strong positive scores (>5) in terms of perceived politeness. The comparatively high mean score for the Bald proposal (lacking face-redress) suggests that the perceived politeness of a proposal strategy is determined relative to the context in which it occurs and not only as a consequence of using expressions which are commonly associated with politeness, such as "I'm sorry" and "thank you".

Table 4
Within-subject contrasts of the three proposal conditions for questionnaire items that showed a statistically significant main effect of proposal condition [$df = 2$, $*p<.05$; $**p<.01$; $***p<.001$]

Questionnaire Item	Positive face means	Negative face means	Bald strategy means	Positive vs. Negative face redress $F =$	Negative face redress vs. bald strategy $F =$	Positive face redress vs. bald strategy $F =$
The style of the proposal was too formal	5.12	4.45	4.88	13.69**	5.98*	1.61
The proposal was too long	2.85	3.13	3.83	1.19	8.64**	18.00***
The proposal made me feel I was being manipulated	3.09	3.73	3.76	5.50*	.20	13.15**
The proposal was an efficient way of giving information about the On-line Saver account	3.66	3.93	4.23	.60	2.63	6.89*
I found the proposal intrusive	2.48	2.94	2.99	3.29	.83	11.98**
The proposal was polite	5.15	5.65	5.33	6.74*	4.32*	.44
The proposal contained only relevant information	4.34	4.65	4.99	3.93	1.83	7.58**
The proposal was very long-winded	3.49	3.48	4.28	.12	8.39**	7.02*
I found the proposal patronising	3.55	4.19	4.55	3.08	5.01	14.84***
The way the proposal was expressed was too apologetic	5.19	4.53	5.21	11.20**	10.23**	.63

The Bald proposal was considered to contain the most amount of *relevant information* of the three proposals, but this was only the difference between the Positive and Bald proposals that showed statistically significant results. In terms of *efficiency*, there was no preferred strategy among the proposals. Not even the Bald strategy (which is aimed to be short and terse to promote efficiency) was rated strongly positively and it was only slightly more favoured than the Positive face-redressive proposal.

In summary, the analysis of the pooled responses highlighted differences in participants' perception of the contrasting politeness styles and registers employed in the proposals. The Bald proposal strategy was perceived to be significantly shorter and less long-winded than the Positive and Negative face-redressive strategies. In line with Brown and Levinson's theory, the wording in the Negative face-redressive strategy was perceived to be more formal, more polite and more apologetic. The Positive face-redressive strategy was rated as the most manipulative of the three proposals and it was considered significantly more patronizing and intrusive than the Bald strategy.

7.3. Task completion

In each call, participants were asked to telephone the service and find out the balance of their current account and then to take a note of the amount on their task sheet. In the third call to the service, participants experienced the product proposal and then had to accept or reject the proposal to set up an (On-line Saver) savings account straight away. Following this, the automated service then asked participants "Would you like another service?" and participants were required to answer "yes" in order to proceed with their account balance enquiry. Successful balance task completion rates for the two (practice) phone calls (1 and 2) and the proposal phone call (3) are shown in Table 5.

The lowest task completion (76%) occurred in the participant group which experienced the Bald style of proposal. When participants in this group were asked if they would like another service, seven out of 22 individuals answered (wrongly) "no" and their call was transferred from the service.

7.4. Interview comments

At the end of the experiment session an opportunity was taken to investigate each participant's reactions to a number of issues raised by their experience of the product

Table 5
Task completion rate: participants who manage to obtain the account balance in a call

Proposal condition	First phone call	Second phone call	Third phone call
No proposal	26 (100%)	25 (96%)	25 (96%)
Positive	28 (97%)	29 (100%)	26 (90%)
Negative	28 (100%)	28 (100%)	28 (100%)
Bald	29 (100%)	29 (100%)	22 (76%)

proposal in the service, including direct comparisons of the three different politeness styles of proposals. This involved a structured interview in which the question order and wording remained the same for each participant. Most of the questions in the interview required the participant to select a proposal of their choice,[8] with the option for participants to volunteer additional comments. Participants were encouraged (but never required) to give more detailed reasons for their responses. The purpose of the interview was to allow participants to express more freely their thoughts about the wording and style of the proposals.

When asked about *which of the three proposals they preferred*, the majority of participants (54%) chose the Bald strategy stating it was shorter and more to the point than the other two. This group also commented that they perceived the Bald style of the proposal as "less patronizing", "less intrusive", "less formal", "more honest" and "more professional". The Negative face-redress proposal received 29% of participants' votes for preferred choice, claiming that they preferred it because it was "polite" and "apologetic" and referred to specific appealing expressions used in the proposal such as "sorry to interrupt", "bank's policy" and "happy to set up". There was no strong consensus in the comments for participants who said they preferred the Positive face-redress (11%). Examples of comments were that the positive style proposal was: "more positive", "more caring", "more polite" and "not so apologetic". Interestingly, when examining only the responses from the control-group participants (who had not experienced the proposal during their use of the automated service), 50% of participants preferred the *Negative* face-redress proposal whereas 38% were in favour of the Bald strategy and 12% the Positive face-redress.

When asked which of the three proposals they perceived as the *most polite way to address the caller*, the Negative face-redress strategy generated a majority (66%) of participants' votes. Consistent with Brown and Levinson's theory, most of the participants who selected the Negative face-redress as the most polite regarded the apology in the opening statement of the proposal as the primary reason for their choice. Here, 18% of participants chose the Bald strategy as the most polite way to address the caller, mainly commenting on that they thought it was "less patronizing" and "less apologetic". The remaining 13% of participants who chose the Positive face-redress proposal did so as they thought it was the most polite way to address the caller. Their comments were: "more familiar", "more natural" and "not as blunt as the Bald strategy nor as apologetic as the Negative face-redress".

In addition, participants were asked which of the proposals was the *least polite way to address the caller.* The Positive face-redress was selected by 48% of participants as the least polite way, mainly commenting on the opening statement in the proposal "I know you won't mind" which many perceived as presumptuous. Participants who chose the Bald proposal (35%) as the least polite of the three said they found it "abrupt" and that they did not like the opening statement (I'm interrupting to inform you...). Only 5% of participants thought that the Negative face-redress was the least polite way to address the caller. In this case, half of the

[8] Some participants selected more than one proposal. In order to simplify the discussion in this section, only participant responses where one proposal was selected were included in the analysis.

comments regarded the statement about "bank's policy" as indicative of less concern about the customer's finances.

Finally, participants were asked which of the proposals they found to be *the friendliest*. In this case, the Negative face-redress proposal received the majority (49%) of participants' votes with reasons that it was "apologetic", "personalized", "more human sounding" and that they liked the phrase "valued customer". Participants who thought the Positive face-redress proposal was friendliest (29%) said they found it to be "more informal" and "more personalized". The remaining 15% who selected the Bald strategy did so mainly because they thought it was better in comparison with either of the face-redressive proposals.

In summary, participants comments indicated a preference for the Bald strategy to be employed for system-initiated digressive dialogues. However, a significant proportion of participants favoured some kind of politeness strategies indicating that users were aware of the importance of mitigating face strategies in human–computer interaction.

8. Discussion

This paper has described an experiment in which participants ($N=111$) experienced a digressive proposal offering a new product to the caller as part of the interactive dialogue of an automated telephone banking service. The opening phrase in the proposal explicitly stated that the proposal constituted an interruption. Three contrasting politeness strategies (Positive, Negative and Bald), derived from established face-redress theories in human–human communication, were employed in order to mitigate the adverse effects of these dialogue intrusions. Participants' attitudes towards these three proposals were explored, both in terms of their impact on perceived usability of the banking service and the perception of the interrupting digression itself (as moderated by politeness strategy).

The experiment data presented in the paper reveal that the usability of the spoken telephone banking service is reduced with the introduction of these digressive interruptions in the dialogue. Participants' initial mean usability score of 5.69 (7-point response scale) fell to a mean score 4.96 after they had experienced their first proposal. This significant reduction in usability was observed for each of the three politeness strategies explored—Positive face-redress, Negative face-redress and Bald register. Participants found the service with such proposals "more frustrating to use", "less enjoyable", "less efficient" and "more in need of improvement". The proposals also placed more cognitive strain on the participants rendering the interaction "more confusing", the service "more complicated" and "less easy to use". Interestingly, the results show that the types of apology and politeness used in the Negative face-redress strategy (which are typically associated with politeness etiquette) were not effective. The use of "I'm very sorry to interrupt..." in the Negative face-redress was no better received than the phrase "I'm interrupting..." in the Bald strategy.

Despite employing contrasting polarities of face-redressive strategies, there were no overall significant differences between the three proposals, based on participants' first exposure. Two real-life issues were considered in detail in the design of the experiment that may have modified participants' perception of the proposal, making the dissimilarities between the contrasting strategies more prominent. Firstly, since the digressive proposal forms only a brief part of a larger automated banking dialogue, the impact of the contrasting proposal strategies may have been more strongly differentiated had participants been forewarned about the pending interruption; this may have encouraged participants to pay more careful attention to the contents and wording of the message. In a real-life scenario, however, it is difficult to envisage how and when such warnings might be delivered to customers and, in consequence, the un-primed (worst-case) approach was adopted in the experiment. Secondly, the scenarios in the experiment might have been extended to involve a secondary task implicitly instructing participants to maximize their savings returns, thereby making them more positively disposed towards the product introduced in the digression offer. However, if such digressions were introduced in a real-world automated telephone banking service, possibly based on some assessment of a customer's individual need for the product, there is no guarantee that the customer would actually share the enterprise's perceived need for that information. Hence in the experiment design a totally un-targeted (worst-case) approach was adopted. The un-primed, un-targeted scenario approach had been used successfully in previous dialogue digression experiments by the authors (Wilkie et al., 2002).

The analysis of the pooled response data (after participants had experienced each of the three proposals) revealed that there were significant differences overall between the Bald and the Positive face-redress strategies (in favour of the Bald strategy). The results provide some guidance on the design issues involved in attempts to add such digressions to automated telephone dialogues by eliciting participants' preferences for the wordings of such proposals. The Bald strategy received significantly more positive responses in terms of being shorter, less long-winded and contained more relevant information. The Positive face-redress, on the other hand, was found to be significantly more manipulative, patronizing and intrusive. In the post-experiment listening tests, support of the Bald proposal strategy was strengthened: 54% of participants expressed a preference for the Bald strategy with the main arguments that it was shorter and more to the point than the other designs. Interestingly, however, 50% of participants in the (No-proposal) control group chose the Negative face-redress as their preferred proposal strategy. When heard in isolation, the Negative face-redress might seem the most appropriate design choice when approaching a customer; in the context of the automated telephone banking service, however, the Negative face-redress approach was shown to be judged as lengthy, long-winded and was perceived to be too apologetic and formal.

Much of the research of anthropomorphic computer behaviour in human–computer interaction to date has primarily focused on the visual user interface; the impact of social phenomena, such as politeness, in the audio-only interface have yet to be fully explored. The current research contributes to the debate on anthropomorphism in computer systems by exploring the issue of endowing a speech-only human–computer dialogue with specific forms of politeness. In contrast to the visual user interface, the audio-only interface is incapable of displaying multiple pieces of information simultaneously; the system will dominate the dialogue for as long as it takes to deliver a spoken message and the user is not offered the opportunity to rapidly scan information that seems irrelevant. It follows that choice of appropriate wording, duration and speaker characteristics are pivotal in the design of audio interfaces—as demonstrated in this paper. These issues raised here lend themselves to further research in order to obtain a deeper understanding of pronounced forms of speaker characteristics, linguistic behaviour and user expectations unique to such audio-only computer interfaces.

For a given communicative situation between humans, it has been shown that the choice of politeness strategy depends on the mutual expectations about the power relationship and social distance between the interactants, coupled with the degree of imposition involved in making the face-threatening act in that communicative context. When a speaker is overly polite, unexpectedly unfriendly or irrational, or strays from the topic in a human–human conversation, the addressee will draw conclusions about the reasons why the speaker does not behave as expected. This may, e.g. involve re-evaluating the assumptions about their social relationship with the consequence that politeness (or its absence) in a dialogue can serve to modify the social distance or power relationship between interactants. The negative reactions towards the face-redressive strategies employed in the proposals may be attributed to the fact that these were not perceived as being fully integrated with users' assumptions about the relationship with the service, formed by the speaker characteristics presented in the rest of the banking dialogue. Whilst applications, such as the automated banking service explored in this research, are primarily viewed as tools, with repeated use there is the potential that customers will develop aspects of rapport with the service. Endowing audio-only interfaces with personas, which consistently exhibit Negative or Positive face-redressive behaviour as presented here, may thereby serve to enhance this human–computer relationship.

Acknowledgements

The current research was made possible by the generous support of Lloyds TSB Group, UK and Nuance Communications, Inc.

Appendix A

In order to establish the absolute politeness in the proposals (i.e. attitudes towards the politeness strategies when removed from the context of the telephone dialogue), an additional session was included at the end of the experiment in which control-group participants listened to each proposal over computer speakers. The aim of the

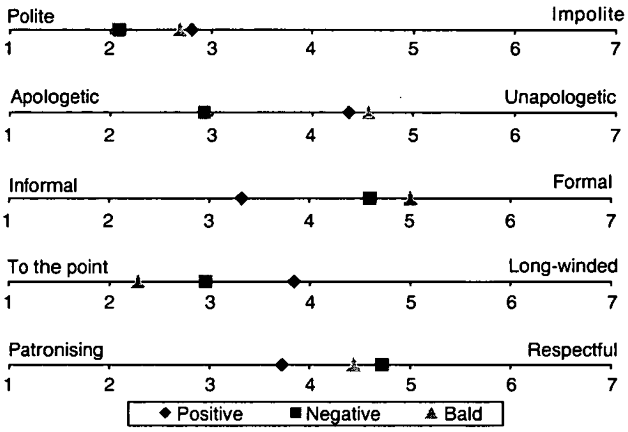


Fig. 4. Control-group participant mean responses. These questionnaire items were introduced to participants with the phrase "*Thinking about the proposal I've just heard, it was...*".

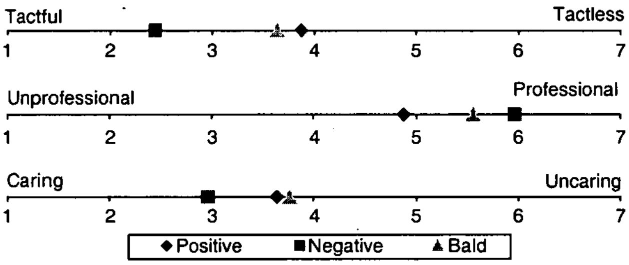


Fig. 5. Control-group participant mean responses. This section was introduced to participants with the phrase "*I associate the choice of wording in the proposal with someone who is...*".

listening session was two-fold: (1) explore the participant's perception of the register and speaker characteristics employed in the contrasting proposals and (2), whether the contrasting face-redress strategies would produce effects consistent with Brown and Levinson's theories.

Immediately after hearing a proposal, participants completed a questionnaire featuring descriptive antonym pairs (such as polite vs. impolite) are presented at either end of a 7-point (semantic differential) scale (Osgood et al., 1957). The first set[9] of antonyms concerned the style and register used in each of the contrasting proposals and were introduced to the respondents with the sentence: "*Thinking about the proposal I've just heard, it was...*". Respondents marked their opinions by ticking the appropriate box along the scale. The second set[10] of antonyms were aimed at assessing some of the social characteristics and personality of the speaker. Response

[9] The statements were: polite/impolite; informal/formal; to the point/long-winded; forthright/diplomatic; sincere/insincere; respectful/patronizing; personalized/impersonal; apologetic/unapologetic.

[10] These items were: tactful/tactless; timid/self-confident; sociable/unsociable; reliable/unreliable; caring/uncaring; professional/unprofessional.

Table 6
Statistical analysis of the absolute politeness in the contrasting proposals in the listening session [$df = 2$, $*p < .05$; $**p < .01$; $***p < .001$]

	Positive vs. Negative face redress $F =$	Negative face redress vs. Bald strategy $F =$	Positive face redress vs. Bald strategy $F =$
Polite–impolite	3.47	5.82*	.38
Apologetic–unapologetic	8.42**	11.91**	3.94
Informal–formal	8.95**	2.18	29.07***
To the point–long-winded	21.39***	.84	10.40**
Patronising–respectful	31.34***	2.24	2.84
Tactful–tactless	20.48***	13.83***	.07
Unprofessional–professional	19.84***	2.71	3.46
Caring–uncaring	4.71*	7.65*	1.39

data was first pooled according to the politeness strategy employed (Positive, Negative and Bald) and a three-level repeated-measures ANOVA was then performed based on individual questionnaire attributes. Figs. 4 and 5 summarize the results of the absolute politeness check (further details of the statistical significance are included in Table 6).

The tendency to view the Negative strategy as polite, apologetic and tactful is consistent with Brown and Levinson's theory which states that listeners commonly associate expressions of Negative face-redress with the everyday use of the term politeness—it is "the stuff that fills etiquette books". In contrast, the Positive face-redress is realized through more intimate linguistic output strategies where the aim is to show appreciation and care of the addressee's wants in general, or express the similarity between the speaker and addressee's wants.

The Positive face-redress was perceived to be significantly more *long-winded* than the Negative face-redress ($p < .01$) and the Bald strategy ($p < .01$). This is interesting as it indicates that it is not primarily the duration (the Positive and Negative face-redress proposals were of the same length, ±1 s) but the choice of wording that contribute to the addressee's perception of long-windedness in the proposal.

References

Allen, J.F., Guinn, C.I., Horvitz, E., 1999. Mixed-initiative interaction. IEEE Intelligent Systems 14 (5), 14–23.

Anderson, R.I. (Ed.), 2000. Conversations with Clement Mok and Jakob Nielsen, and with Bill Buxton and Clifford Nass. Interactions 7(1), 46–80.

Balentine, B., Morgan, D.P., 2001. How to Build a Speech Recognition Application, second ed. EIG Press, ISBN: 0-9671287-2-3.

Bernsen, N.O., Dybkjær, H., Dybkjær, L., 1997. Elements of speech interaction. In: Dybkjær, L. (Ed.), Proceedings of the Third Spoken Language Dialogue and Discourse Workshop, Vienna, September 1997.

Boyce, S.J., 2000. Natural spoken dialogue systems for telephony applications. Communications of the ACM 43 (9).

Brown, P., Levinson, S.C., 1987. Politeness: Some Universals in Language Use. Cambridge University Press, Cambridge.

Chen, R., 2001. Self-politeness: a proposal. Journal of Pragmatics 33, 87–106.

Colón, J.X.E., Pérez-Quiñones, M.A., Ferreira, R., 2001. Effects of face-threatening acts in human–computer dialogues. Proceedings of HFES'01 (Human Factors and Ergonomics Society), pp. 657–662.

Culpeper, J., 1996. Towards an anatomy of impoliteness. Journal of Pragmatics 25, 349–367.

Fogg, B.J., Nass, C., 1997. Silicon sycophants: the effects of computers that flatter. International Journal of Human–computer Studies 46, 551–561.

Gardner-Bonneu, D., 2001. Human Factors and Voice Interactive Systems. Kluwer Academic Publishers, Dordrecht, 1999, Second Printing 2001, ISBN 0-7923-8467-9.

Haller, S., 1994. Recognizing digressive questions. In: Proceedings of AAAI94, Fall Symposium 1994.

Hone, K.S., Baber, C., 1999. Modelling the effects of constraint upon speech-based human–computer interaction. International Journal of Human–Computer Studies 50, 85–107.

Leech, G.N., 1983. Principles of Pragmatics. Longman Group, Ltd., New York.

Likert, R., 1932. A technique for the measurement of attitudes. Archives of Psychology 140.

Love, S., Dutton, R.T., Foster, J.C., Jack, M.A., Nairn, I.A., Vergeynst, N.A., Stentiford, F.W.M., 1992. Towards a usability measure for automated telephone services. Proceedings of the Institute of Acoustics Speech and Hearing Workshop 14 (6), 553–559.

McFarlane, D., 1998. Interruption of people in human–computer interaction. Ph.D. Thesis, The George Washington University, August 1998.

Narayanan, S., Di Fabbrizio, G., Kamm, C., Hubbell, J., Buntschuh, B., Ruscitti, P., Wright, J., 2000. Effects of dialog initiative and multi-modal presentation strategies on large directory information access. In: Proceedings of the International Conference on Spoken Language Processing, ICSLP 2000, Beijing, China, pp. 636–639.

Nass, C., Moon, Y., 2000. Machines and mindlessness: social responses to computers. Journal of Social Issues 56 (1), 81–103.

Nass, C., Steuer, J., Tauber, E.R., 1994. Computers as social actors. Proceedings of the CHI '94, Conference of the ACM/SIGCHI, Boston, MA, April 1994.

Osgood, C.E., Suci, G.J., Tannenbaum, P.H., 1957. The Measurement of Meaning. University of Illinois Press, Champaign, IL.

Ramakrishnan, N., Capra, R., Pérez-Quiñones, M.A., 2002. Mixed-initiative interaction = mixed computation. In: Thieman, P. (Ed.), Proceedings of the ACM SIGPLAN Workshop on Partial evaluation and Semantic-Based Program Manipulation (PEPM'02), January 2002, pp. 119–130.

Reeves, B., Nass, C., 1996. The Media Equation. Cambridge University Press, Cambridge.

Shneiderman, B., 1988. A nonanthropomorphic style guide: overcoming the Humpty Dumpty syndrome. The Computing Teacher 9–10.

Shneiderman, B., 1993. Beyond intelligent machines: just do it!. IEEE Software 10 (1), 100–103.

Shneiderman, B., 1998. Designing the User Interface: Strategies for Effective Human–Computer Interaction, third ed. Addison Wesley Longman, Inc.

Shneiderman, B., 2000. The limits of speech recognition. Communications of the ACM 43 (9).

Wilkie, J., Jack, M.A., Littlewood, P., 2002. Design of system-initiated digressive proposals for automated banking dialogues. Proceedings of the International Conference on Spoken Language Processing, ICSLP 2002, pp. 1493–1496.

DESIGN OF SYSTEM-INITIATED DIGRESSIVE PROPOSALS FOR AUTOMATED BANKING DIALOGUES

Jenny Wilkie[1], *Mervyn A. Jack*[1], *Peter Littlewood*[2]

[1]Centre for Communication Interface Research
The University of Edinburgh, U.K.

[2]Lloyds TSB Group, Bristol, U.K.

Abstract

System-initiated proposals may be used to introduce new and unsolicited information into the dialogue flow of an automated telephone service in order to advise callers about products in which they may be interested such as short-term loans or overdrafts. Important dialogue design issues surrounding the introduction of such digressive proposals include *how* to interrupt the callers and *where* in the dialogue flow it is most suitable to locate a proposal. This paper describes the results from two experiments using a spoken natural language telephone banking application where two delivery strategies and three contrasting locations were investigated. Results showed that the delivery strategy had a stronger effect than the location.

1. INTRODUCTION

There are distinct business advantages in terms of cost saving and 24x7 accessibility associated with employing automated telephone applications in areas such as banking to enhance customer services. However, successful take-up of such applications may result in the enterprise losing opportunities for direct contact with customers, for example opportunities to advise them about new or relevant products and services, possibly based on the individual's banking profile. A solution to this problem involves the introduction of system-initiated informational prompts (sales proposals) within the dialogue structure with the intention of cross-selling new products and services to the customer. These 'Sales through Service' (StS) proposals may be viewed as an extreme form of system-initiation in dialogues since they are not related directly to the current topic or to the prime goal of the call – hence the term digressive proposal.

Digressions are common in human-human conversation where participants use their knowledge about coherence, states of attention and intention in the discourse to find the appropriate timing to introduce new topics into the flow of a conversation [1]. This intrinsic human ability to co-ordinate and collaborate in interactive activities poses a challenge to designers of automated human-computer interfaces, particularly in the area of mixed-initiative interaction [2], [3].

Research into digression in human-computer interactions (often referred to as 'out-of-turn interaction' or 'unsolicited reporting' [4]) has focused mainly on providing models for handling *user*-initiated digression [5], [6], [7] where the user supplies extra or out-of-turn information in response to system prompts. Results from experiments into system-initiated back-channel feedback in human-computer spoken interfaces [8], [9] also suggest that the style and timing of such responses may affect user impressions of a service.

Human-computer dialogues in mass-market telephone applications generally employ a fixed-initiative strategy where the system prompts the user for information and the dialogue does not usually change between phone calls. System-initiated digressive proposals, which interrupt the regular call flow with extra information, therefore run the risk of being perceived as disruptive or distracting by habituated users.

2. APPROACH

It is expected that user attitudes to system-initiated digressive dialogue proposals will vary according to the relevance of the information to the user's specific situation. Determining what is or is not relevant to an individual caller is a complex matter involving modeling of the caller's intentions, wants, needs and goals. The research reported here did not address issues relevant to defining the criteria for deciding whether or not to make a proposal to a particular caller on a particular occasion. Rather, it addressed dialogue design issues relating to how to deliver and where to place digressive proposals in automated telephone service dialogues, assuming that the decision has already been taken to make the proposal.

The digressive proposals explored in this research were designed to inform callers about the availability of an overdraft and to give instructions on how to obtain the overdraft from within the service. A key issue with respect to the design of digressive proposals is their degree of obtrusiveness: a proposal needs to be prominent enough to capture the attention of interested users but not so prominent as to impact negatively on attitudes to the service.

3. EXPERIMENT DESIGN

3.1. PhoneBank *Express* Dialogue Overview

The automated telephone service used in this research was based on a commercially available telephone banking service provided by Lloyds TSB - PhoneBank *Express*. This banking service allows customers to find account balances, transfer money, hear a list of transactions, pay bills etc. by using spoken natural language input (in English). A high-level flow-chart of the dialogue architecture is outlined in Figure 1 below.

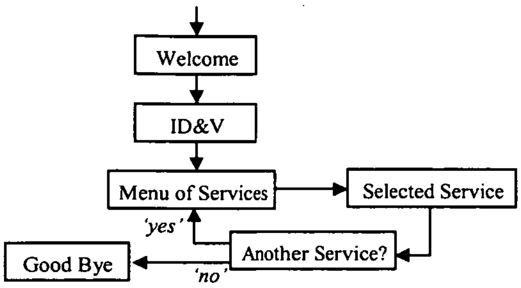


Figure 1: PhoneBank Express Dialogue Flow-chart

The first stage in the dialogue is a Welcome greeting: "Welcome to PhoneBank *Express*". The customer is then asked to enter a membership number and two random digits from a secret identification number (TIN) in the Identification and Verification stage (ID&V). Following successful identification the caller hears, and selects an option from, the Menu of Services ***"Please select balance, recent transactions or another service"*** ('another service' calls the second part of the menu: ***"In addition, you can select funds transfer, item search, order statement or change TIN"***). Each return to the Menu of Services is preceded by a question ***"Would you like another service?"*** (answering 'no' to this ends the call).

3.2. Proposal Strategy

For the purposes of the first experiment reported here the basic PhoneBank *Express* dialogue was augmented with a proposal strategy. System-initiated proposals can take one of two forms, referred to here as 'Signpost' and 'Follow-on'. The proposals were located in the dialogue flow at a point following the readout of the balance of the current account (after 'Selected Service' in Figure 1).

The **Signpost Strategy** consisted of a short message embedded within the normal service dialogue informing callers about the availability and location of the overdraft option in the automated dialogue (in this case the overdraft option was available at the menu of services). The intention behind the Signpost Strategy is to interest and inform the caller without intruding too heavily on the call flow. It is then at the caller's discretion to locate and select the product option within the dialogue structure. The wording of the Signpost proposal prompt was as follows: ***"You might like to know that you can have an overdraft on your current account. To find out more, just say overdraft at the menu of services."***

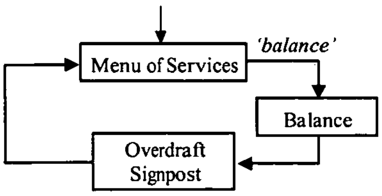


Figure 2: Signpost Proposal Strategy

Potentially more intrusive, the **Follow-on Strategy** involves prompting the caller who must then make a decision (and respond 'yes' or 'no') to either accept or reject the offer before the dialogue can continue. If the caller agrees to the proposal, the system starts a 'follow-on dialogue' giving relevant information about the details or terms of the overdraft and confirming the agreed amount. Callers who declined the proposal were given a Signpost message with information about how to (later) obtain an overdraft (by saying 'overdraft' at the menu of services). The wording of the Follow-on proposal was as follows: ***"You might like to know that you can have an overdraft on your current account. Would you like to arrange an overdraft now?"***

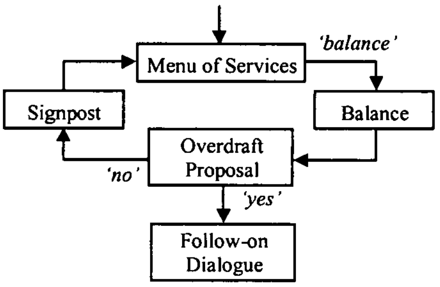


Figure 3: Follow-on Proposal Strategy

3.3. Proposal Location

The second experiment explored *where* it is suitable to make a proposal in the spoken dialogue. Participants were offered system-initiated overdraft proposals using the Signpost Strategy in one of three contrasting locations in the dialogue (see Figure 1 above). The locations were: (1) at the Welcome stage of the dialogue; (2) following a successful completion of the ID&V stage; or (3) after a specific Selected Service transaction (following the balance of the current account).

The **Welcome Proposal** location followed the introductory message in the service ***"Welcome to PhoneBank Express"***. It was worded to be applicable to all callers. Due to its location within the dialogue, it might be expected that the proposal would pose a lower risk of distracting the caller from their task at hand. However, because the caller had not yet been identified at this stage in the dialogue there was a risk that prospective applicants may have to be turned down. The wording of the Welcome Proposal was as follows: ***"[Welcome to PhoneBank Express.] We've added a new overdraft facility to this service. To find out more, just say overdraft at the menu of services."***

The **ID&V Proposal** location followed immediately after the successful verification of the caller in the dialogue. The proposal could therefore be made customer-specific with targeted information about the particular account (such as the allowed overdraft limit), reducing the risk of having to turn down the applicant. The wording of the ID&V Proposal was: ***"You might like to know that you can have an overdraft of £400 on your current account. To find out more, just say overdraft at the menu of services."***

Finally, the **Transaction-linked Proposal** location was a nested, system-internal prompt that followed a particular transaction or sub-dialogue in the service and would be used to link the proposal information to certain account details, a particular transaction, topic or service in the dialogue. The ability to create a logical link between proposal information and specific details in the dialogue can be useful but it can also be potentially more distracting to the caller who may be heavily involved with the task at hand. The Transaction-linked Proposal follows immediately after the current account balance readout and was worded as follows: ***"You might like to know that you can have an overdraft of £400 on this account. To find out more, just say overdraft at the menu of services."***

3.4. Experiment procedure

In each of the two experiments carried out using these modified PhoneBank *Express* dialogues, participants each made three phone calls to the service undertaking two banking tasks in each call (finding and noting down the balance of their current account followed by ordering a statement for their savings account). The first two phone calls involved use of PhoneBank *Express* without StS proposals, allowing callers to become familiar with the service functionality. The third phone call included the overdraft proposal dialogue. In the Proposal Strategy phase of the experiment, one-third of callers constituted a control group and did not experience an overdraft ('no-proposal version').

After the first two phone calls to the service, participants completed a usability questionnaire which used a standard Likert format [10] to assess participants' attitudes towards the automated telephone banking service interface. Four basic aspects of usability were covered: cognitive issues, quality of interface and system performance, transparency and fluency of the service, and conversational model [11]. The data obtained from this questionnaire were used as a baseline reference for participants' attitude towards the PhoneBank *Express* service.

A second usability questionnaire was completed following the third call to the service in order to allow investigation of the impact on usability and attitudes of the digressive proposal. The results from this questionnaire were used to compare the attitude towards the StS proposal dialogue between proposal groups. This second questionnaire, again with a Likert format, included items specifically designed to elicit information directly related to the proposal experienced: *intrusiveness* (proposal was annoying, intrusive, too long, interrupted the call, distracting); *user confidence in the service* (trust the information, happy to apply through the service, relying on the service when applying, preference of having a human giving the proposal information); *quality of the proposal* (helpfulness of proposal information, efficiency of method, level of politeness); *cognitive effort* (easy to understand, knowing how to use the service to apply for an overdraft) and *relevance of the proposal.*

4. EXPERIMENT RESULTS

To investigate the impact of the system-initiated proposals, repeated-measures analysis of variance (ANOVA) was carried out using the mean responses to the Likert usability questionnaire completed after the two familiarisation calls and after the third call (for calls where the overdraft proposal was experienced). To compare the effect and attitude towards the different proposal locations and strategies used in the experiments uni-variate ANOVAs were carried out on those Likert questionnaire items specifically addressed to the proposals. Three between-subject factors were included in the analysis as follows: age group; gender and proposal condition.

4.1. Proposal Strategy Results

A total of 179 callers contributed data to the analysis for Proposal Strategy. Re-assuringly there were no significant differences between the overall mean responses to the general usability of the service before and after experiencing a proposal. The results suggest that neither the presence of a proposal nor the type of the proposal explored in this experiment impacted (positively of negatively) on the overall perceived usability of the banking service. Two attributes were more statistically positively rated ($p<0.05$) for the no-proposal version of the service with participants finding that the service without the proposal needed a lesser *degree of concentration* and was a *more efficient service.* Detailed analysis of the perceived usability data overall confirms that there were no significant interactions between the proposal conditions experienced by the participants and any of the participant factors, age and gender.

There were, however, highly significant differences in responses to the items relating to the proposal strategies. Those who experienced the Follow-on Proposal took a more negative attitude to the *length of the proposal* than did those who experienced the Signpost Proposal (mean for Follow-on = 4.28; mean for Signpost = 5.02, both on a 7-point scale where scores above 4.0 indicate positive attitudes, $p<0.01$). The Follow-on Proposal Strategy was also perceived to *interrupt the call* more than the Signpost Proposal Strategy (mean Follow-on = 3.72; mean Signpost = 4.65, $p<0.01$). At a lower level of significance ($p<0.05$) the Follow-on Proposal Strategy was also considered to be *more annoying* and *less appropriate* to this kind of service and *more distracting* than the Signpost Proposal Strategy. However, participants thought they *knew better how to apply* for an overdraft having experienced the Follow-on Proposal Strategy.

4.2. Proposal Location Results

A total of 119 participant data sets were used in the analysis of Proposal Location. There were no significant differences in overall usability between the three proposal locations and no effects for gender or age. The results suggest that, in terms of overall usability, the actual location of the proposal had little effect on participants' attitude towards the service.

More differences were found with respect to questionnaire items relating specifically to the proposal location. Two items showed statistically significant differences between the three proposal locations in the dialogue: *appropriateness for the proposal for the type of service* and *the perceived length of the proposal.*

The service containing the overdraft proposal as part of its Welcome message was judged to be significantly better (mean=4.65, $p<0.01$) in terms of *appropriateness,* than the Transaction-linked proposal (mean=3.20) and the ID&V proposal (mean=3.46). In terms of *perceived length of the proposal,* participants were more positive ($p<0.01$) towards the length of the proposal when it occurred as part of the Welcome message (mean=5.05) and after the ID&V process (mean=4.87), compared to the Transaction-linked proposal (mean=4.15). Although the lengths of the three proposals did not differ more than two seconds, this result suggests that proposal location does influence callers' perceptions of objective properties of the message such as duration.

Overall, the results suggest that there were differences between attitudes to the service corresponding to the three locations. The design which included making the proposal in the Welcome message tended to receive a more positive evaluation.

5. CONCLUSIONS

The research reported here investigated some key dialogue design issues surrounding *how* and *where* to introduce system-initiated proposals in the dialogues of automated telephone banking services. Two strategies for proposals were designed and investigated (Signpost and Follow-on); and Signpost-style proposals were tested in three locations in the dialogue (Welcome message, ID&V process and Transaction-linked).

Participants made two phone calls to a stand-alone mirror version of PhoneBank *Express*. On a third call they experienced a digressive overdraft proposal. Data show that, in terms of usability of the service, neither the location nor the strategy of the proposal in these experiments has any overall effect on callers' attitudes to the service.

When studying the Proposal Location, the service containing the overdraft proposal in its Welcome message was judged to be significantly better in terms of the *appropriateness* of the proposal for the type of service. Participants were more positive towards the *perceived length* of the proposal when it occurred in the Welcome message (or after the ID&V process) compared to when the proposal followed a transaction (balance of the current account).

Experimental results for the Proposal Strategy for digressive overdraft proposals showed a measured lower attitude to the *length of time* the Follow-on Proposal Strategy took relative to the Signpost Proposal Strategy and the Follow-on Proposal Strategy was perceived to *interrupt the call* more than the Signpost Proposal Strategy. The Follow-on Proposal Strategy was also considered to be *more annoying* and *less appropriate* to this type of service and *more distracting* than the Signpost Proposal Strategy. However, participants thought they *knew better how to apply* for an overdraft with the Follow-on Proposal Strategy.

6. REFERENCES

[1] Lenk, U., (1998) Discourse markers and global coherence in conversation. *Journal of Pragmatics 30 (1998) pp. 245-257.*

[2] Haller, S., S. McRoy, "Computational Models for Mixed Initiative interactions", *Papers from the 1997 AAAI Spring Symposium. Technical Report: SS-97-04: AAAI/MIT Press, 1997.*

[3] Horvitz., E., "Principles of Mixed-Initiative User Interfaces", *In Proceedings of ACM CHI 99 Conference on Human Factors in Computing Systems, Vol. 1 of Characters and Agents, pp. 159-166, 1999.*

[4] Allen, J. F., C. I. Guinn, E. Horvitz, "Mixed-initiative interaction", *IEEE Intelligent Systems, Vol. 14(5): pp. 14-23, Sept-Oct 1999.*

[5] Ramakrishnan, N., R. Capra, M. A. Pérez-Quiñones, "Mixed-Initiative Interaction = Mixed Computation", *In Proceedings of the ACM SIGPLAN Workshop on Partial evaluation and Semantic-Based Program Manipulation (PEPM'02) (P. Thieman, ED.), pp 119-130, January 2002.*

[6] Narayanan, S., G. Di Fabbrizio, C. Kamm, J. Hubbell, B. Buntschuh, P. Ruscitti, and J. Wright, "Effects of dialog initiative and multi-modal presentation strategies on large directory information access", *In Proc. of ICSLP, (Beijing, China), pp. 636-639, 2000.*

[7] Haller, S., "Recognizing Digressive Questions", *In Proceedings of AAAI94, Fall Symposium 1994.*

[8] Hirasawa, J., M. Nakano, T. Kawabata, K. Aikawa, "Effects of System Barge-in Responses on User Impressions", *In Proceedings of Eurospeech 1999, 6th European Conference on Speech Communication and Technology, Vol. 3, pp. 1391-1394.*

[9] Okato, Y., K. Kato, M. Yamamoto, S. Itahashi, "System-user interaction and response strategy in spoken dialogue system", *In Proc. of ICSLP 1998.*

[10] Likert, R., "A Technique for the Measurement of Attitudes", *Archives of Psychology, Vol. 140, June 1932.*

[11] Love, S., R.T. Dutton, J.C. Foster, M.A. Jack, I.A. Nairn, N.A. Vergeynst and F.W.M. Stentiford, "Towards a usability measure for automated telephone services", *Proc. Institute of Acoustics Speech and Hearing Workshop, vol.14, no.6, pp.553-559, November 1992.*

Proceedings of ICSLP'02, Colorado, pp. 1493-1496